Contents

Acknowledgments

I feel very privileged to have undertaken a dissertation on cutting edge network theory topics that were published only weeks before it was started. I would like to thank my supervisor Dr. Mason Porter for his guidance and advice through this very vast research area that encompasses different types of literature written by different people in different disciplines. This dissertation would not have been possible without the help of his experience and perspective of this broad field. I also would like to thank Professors Alex Scott and Oliver Riordan for very exciting lectures that motivated me to take a dissertation in this field and for help with researching papers. A special thanks must also go to my family who supported and motivated me to come to Oxford and study this degree program, and a final thanks to all my coursemates for making this a very exciting and enriching experience.

Abstract

For many decades the random graph model presented by Erdős and Rènyi [12] has been the main subject of study of random graphs. Many of it's key properties have been rigorously proven [3] such as the Poisson degree distribution, component sizes, and existence of a giant component etc. However, this model fails to reflect many of the properties found in the real world such as arbitrary degree distributions and clustering. The issue of degree distributions has been mostly solved by the configuration model studied by [2] and [22], in which any degree distribution can be achieved, subsequently Molloy and Reed [22, 23] rigorously proved vertain key properties about this model such as the tree like structure of small components which has subsequently been exploited by applied mathematicians and physicists such as Newman [24] in deriving approximative methods to compute further key properties. These results have been know for about a decade now.

On the other hand, there has been far less success in creating random graph models with clustering that reflects that of real word network. Indeed, the configuration model has a zero clustering coefficient in the limit of a large graph size. Since then, many attempts have risen to derive a model with non zero clustering but were unsolvable analytically such as [16].

Very recently, few authors have published papers claiming that they have a solution to this long standing problem in network theory. Newman [29] and Gleeson [14] introduced models that are essentially a generalisation of the classical configuration model and which have provable non zero clustering in the limit of large graph size. In deriving key properties of graphs of this model, they use methods based on similar assumptions about the structure of graphs in the classical configuration model namely the locally tree like structure.

Our aim in this dissertation is to make the work of Newman and Gleeson more rigorous by demonstrating that their assumptions are justifiable. We will achieve this by adapting the Molloy and Reed [22] proofs of the classical configuration model to the new generalised form. We will then build on this result to present in detail, results shown in Newman's and Gleeson's papers that were derived using tree cascade and generating function methods. We will derive further results not shown in their papers. We will then go further by considering the most general forms of the configuration model conceivable consisting of configurations of any fixed mixed distribution of any motifs, and argue that they must have the same qualitative behaviour as the classical configuration model in having tree like small components and a threshold for the formation of the giant component.

Chapter 1

Background

Introduction

It could be said that study of *random graphs* was started by the highly influential paper[12] by Erdős and Rènyi [12] in which they presented a random graph model where vertices are connected independently and uniformly with a constant probability p. They have also rigorously demonstrated that such a random graph undertakes a qualitative transition above a certain threshold namely the appearance of a giant component. Since then a lot of extensive work has been done on this model and many key properties have been rigorously calculated and proven such as the average geodesic distance, number of cycles, the distribution of the size of small components, motif count etc.

On the other hand the study of *networks*, which is the term used to denote graphs taken from the real world, has seen a new emerging direction in its research shifting from the study of small graphs and local properties to the study of much larger graphs and their global properties. For example, previously a network theorist might have been interested in answering the question what is the shortest path between two given vertices in a transportation network, a newly interesting question now would be to ask what is the average distance between two vertices in a network representing the World Wide Web. This shift has been mainly driven by the appearance of such huge networks like the lately created large online social networks and also the technology that enables network theorists to handle such large amounts of data.

This shift in the scale of networks and properties of interest has also drawn a shift in the method, making random graphs an ideal tool to model such networks. Random graph theory produces results about the structure that apply almost certainly to any graph within a certain family in the limit of large graph size. Hence, armed with an appropriate model and a large enough target real life network one can make very powerful predictions. Random graph models are also used as *null models* in explaining which aspects of networks can be attributed to randomness and which aspects can't.

Random graphs have been used to understand how some real life networks came to have the structure they do. An example of that is the Barabási and Albert's [1] preferential attachment growing model in attempting to explain the degree distribution of the web. They have also been used to compute important properties such as the size of the giant component, which is the proportion of the network connected to each other. Through random graph models we can also predict global behaviour just by knowing local properties: by knowing the degree distribution in a communication network we can predict the existence of a giant component and measure its resilience by computing the percolation threshold, this example in particular we shall see in more detail later.

The importance of network theory for real life applications lies in the fact many real life system

have an underlying structure: the Internet, the brain, the cell. Understanding the structure of such networks can therefore help us understand the behaviour of entities on these networks. Many real life networks are known to display the following properties:

- *Sparseness*: The ratio of the number of edges to the number of vertices tends to a constant in the limit of large graph size.
- *Small world phenomenon*: Any two vertices in the network are connected by a short path (that grows logarithmically with the size of the network).
- *Clustering*: Two vertices that have a common neighbour are more likely to be connected to each other, also called transitivity.
- *Heavy tails*: In their degree distribution, many networks have a significant proportion of vertices with degree significantly higher than the mean.

The classical Erdős and Rènyi random graph model can be tuned to display sparseness by choosing an appropriate connection probability. It is also know to display small world behavious. But it can be shown that in the limit of large graphs it is known to have zero clustering. It is also straightforward to show that for a sparse graph it has a Poisson degree distribution which is not very common in the real world.

In the configuration model [22] the degree distribution is taken as a parameter. The random graph is then constructed by selecting a graph uniformly at random from the ensemble of all graphs with such a degree sequence. A lot of significant work has been done on on this model too and many properties are well know and proven. The configuration model is a very that solves the issue of matching the degree distribution of real world network. However, here again it can be shown that in the limit of a sparse large graph size, a random configuration has zero clustering.

Since this success with degree distributions, very little has been achieved in the next significant characteristic which is clustering. This has limited the prospect of application of random graphs as a tool to model important real life networks such as social networks which are known to have very high clustering. The aim of the work discussed in this dissertation is solve this problem and be able to reflect the type of clustering like that found in social networks.

In this dissertation we will focus on generalised forms of the configuration model introduced by Newman [29] and Gleeson [14] that have non zero clustering. In their papers, the authors use the same methods in computing key properties as they does for the classical configuration model, thereby making the assumption that graphs generated using the new models have a tree like structure as they do in the classical configuration model.

In chapter 1, we will give a brief overview of the study of networks. This will be in brief review of the literature surveyed. We hope this will motivate many of the ideas in later chapters.

In chapter 2, we introduce the classical configuration model, discuss its properties and present the intuition behind the proof of Molloy and Reed [22]. We will also present the generating function formalism developed and used by Newman [24] and justified by Molloy and Reed's result, we will see how it can be used to facilitate the computation of many results. This chapter is a restatement of the authors ideas.

In chapter 3, we will present Newman's new random graph model with clustering.

In chapter 4, we will present an adapted proof of Molloy and Reed's results about the configuration model to Newman's new model, this will be a necessary result to justify the the calculation of key properties computed in chapter 5. Although this is an adapted proof, it is entirely novel in every other aspects and contains corrections to the original proof. Furthermore, the implications of this result in generalising the configuration model discussed later in the chapter 4, are entirely novel and have never been treated in any literature or publication that we know of.

In chapter 5, we compute key properties of Newman's model. Some of these results are derived, but in little detail, in his paper [29]. Other results are not derived in his papers, we show these here for the first time by adapting the generating function method.

In chapter 6, we will introduce Gleeson's model [14] and discuss generalisations of these types of models and methods to compute their key properties. Everything discussed in this chapter, except the presentation of Gleeson's model is novel.

1.1 Real world networks and clustering

In this section we will give a brief overview of the different types of real life networks. We will define key properties that are relevant to their study. We will also briefly discuss some of the most popular random graph models used in recent years. We hope through this section to provide a motivation to the following chapters where our main results are stated.

1.1.1 Types of networks

Many developments in network theory have been driven by observing certain structural properties in real life networks, of these the most studied are listed below. It is worth noting that the classification of these networks very often overlaps. Online social networks can both be classified as technological social.

Social networks

These are networks used to represent social relationships, in these networks vertices usually represent individuals and edges represent relationships between them ranging from friendships to business partnerships. These networks are usually covered in social sciences literature. One of the most famous early works in this area is probably is the small world experiments by Milgram [21].

These networks are characterised by skewed degree distributions [27, 32] and what Milgram's study attempts to show: very short distance between the vertices in the network. They are also characterised by very high clustering and positively correlated degrees of neighbouring vertices. Newman et al, argue that these two properties can be explained by the phenomenon of communities and groups in these networks [28].

Applications in this area, include how the topology of such networks influences the behaviour of individuals. An example is opinion formation: Yu Song et, claim using their model, that the larger the clustering coefficient in the network the easier a consensus takes place [35]. Porter et al, show using techniques drawn from network theory that there exists correlations between the organisational structure and members political positions in the American house of representatives [31].

Definition 1. *The in degree of a vertex is the number of edges directed towards it. The out degree is the number of number of vertices directed away from it.*

Technological & information networks

Examples of such networks include scientific paper citations: One of the most famous studies in this area is that by Price [33] in which he presents a model of growing networks which recreates the power law degree distribution sometimes observed in these networks using a concept called preferential attachment, which is based on the intuition that an already popular paper is more likely to be encountered by an author writing a new paper and therefore is more likely to be cited again.

Another example is the World Wide Web which is the largest sampled network to date. The web is in fact a directed network but also exhibits an approximate power law distribution (see degree distribution) in both its in and out degree. Barabási and Albert [1] created a growing network model also based on preferential attachment to explain such structure. Results of their work were later derived more rigorously by Bollobás and others [7]. Applications on the Web network deal mostly with the link between the structure of web pages and their content. A very famous example of this is the Page Rank algorithm used in the Google search engine.

The Internet is a network where nodes represent communication hubs and edges represent communication links. The Internet display very short path lengths [27] and disassortative correlations between the degree of its neighboring vertices [28], as well as a high clustering coefficient [4].

Applications on information and technological networks are mainly concerned with their efficient exploitation, or how does the topology of such networks affect communication traffic, robustness to damage etc [4].

Biological networks

Food webs are networks where vertices represent species and edges the relation of feeding on or being fed on.

Biological networks have become an essential tool in understanding the function of organisms on a cellular level. It was expected that after sequencing the human genome we would be able to map each gene to a specific function. But this proved not to be the case because of the complex interactions between the different cellular components. This has motivated modeling such interactions using networks.

Jeong et al [19], created metabolic reaction networks for 43 different organisms. He found that characteristics such as scale-free degree distributions, short path lengths, high clustering coefficients were found universally.

Neural networks can be defined on many scales. Nodes can be anything from individual neurons with edges as synaptic connections to brain regions with edges as pathways. These are very sparse networks, they show scale-free degree distributions, small world properties, and cluster organisation [4]. Their study aims to map structure properties with functional properties.

1.1.2 Key properties

We will now look at some properties of networks that are useful in applications and also because we would like to make our random graph models tractable in the sense that we would like to be able to compute and measure these properties and compare them to those in the real world.

The small world effect

The small world effect refers to the property that most pairs of vertices in the network are linked by a very short path through the network. This property can be quantified in terms of the average geodesic distance l.

$$l = \frac{1}{n/2(n+1)} \sum_{j \geq i} D(i,j)$$

Where $D(i,j)$ denotes the geodesic distance from vertex i to vertex j, n is the number of vertices in the network.

By convention, the term *small world effect* has been used to designate graphs in which the average distance is of order $\log(n)$ or slower as function of the size of the graph. This logarithmic scaling can be proved for a variety of models. Riordan and Bollobás [6] have shown that these characteristics always apply for a random graph with power law degree distribution in the limit of large graph size.

Degree distribution

Degree distributions are usually represented by a function p_k which gives the fraction of vertices of degree k or equivalently the probability that a randomly selected vertex has degree k.

In the case of the Erdős and Rènyi random graph [12], we have that $p_k = p = p(n)$ which produces a Poisson distribution in the limit of large graph size. As mentioned earlier, real world

graphs are found to be usually unlike this, in fact they often have a rightly skewed tail. That is if we sketch k against p_k we obtain a long tail for high values of k above the mean.

A common right skewed degree distribution is the power law distribution where $p_k = k^{-\alpha}$ for a constant α. This type of distribution is found in many real world networks but in fact only applies to the tail of the distributions i.e. there exist a threshold above which the power law applies and not before. Values of α have been empirically found in many cases to lie between 2 and 3. Power law degree distributions can be spotted with a straight line on a doubly logarithmic plot of p_k.

Percolation thresholds

The resilience of networks such as communication networks is measured by applying a percolation process [15] in which the graph is gradually destroyed (or built) by the removal (or the addition) of vertices or edges, hence giving rise to many types of percolation.

In *bond* percolation, the edges of the graph are uniformly and independently kept with a probability ϕ and removed otherwise, we say such an edge is *occupied* if it is kept. In *site* percolation a vertex is independently and uniformly occupied with a probability ϕ, otherwise it is removed along with its adjacent edges.

In percolation processes we are interested in the value of ϕ above which we have a giant connected component. This value is called the *percolation threshold* [15].

Site percolation has applications in network resilience such as in communication networks, where an unoccupied vertex represents a communication node failure [10]. The percolation threshold represents the fraction of nodes that can fail whilst still allowing the bulk of the network to communicate. Bond percolation has applications in epidemic spread in social contact networks, where an occupied edge represents a contact between two individuals susceptible of passing on the disease ([36]). The percolation threshold represent the fraction of infectious contacts necessary to cause an epidemic.

Degree correlation

Another question one might ask about networks is what type of vertices tend be connected to each other. Most commonly, this is asked in the context of vertex degrees. We would like to know the extent to which two vertices of certain degrees are related to each other. This is usually measured by the correlation coefficient of the degrees of connected pairs in the network.

$$\rho = \frac{Cov(D_v, D_w)}{\sqrt{Var(D_v)Var(D_w)}}$$

Where D_v, D_w are random variables giving the degrees of two randomly selected pair of connected vertices.

Clustering

As mentioned previously, one the main features in which real world graphs differ from the classical random graphs is clustering, and since clustering is the main focus of this dissertation we need to motivate and define what we mean by clustering.

In many real world networks, especially social networks, we find if node A is connected to node B and node B is connected to node C then node A is very likely to be also connected to C. This transitive property can easily be motivated in a social context by the fact that a friend of one's friend is also very likely to be one's friend.

So a natural way to measure clustering is the probability that two vertices that share a neighbour are themselves neighbours, or alternatively the probability that a connected triple A, B, C forms a triangle. This is given by

$$C^1 = \frac{3 \times \text{number of triangles}}{\text{number of connected triples}}$$

We call C^1 the clustering coefficient. An alternative definition was introduced by Watts and Strogatz, which is a measure of clustering on a local level, we define this by the probability that a triple connected to a vertex i forms a triangle.

$$C_i = \frac{3 \times \text{number of triangles connected to vertex i}}{\text{number of connected triples}}$$

The clustering coefficient of the whole network is the average over all vertices:

$$C^2 = \frac{1}{n} \sum_i C_i$$

It is important to note that this only one of many ways one can quantify clustering. In fact, this type of clustering is referred to as *triadic closure* as it measure the fraction of closed triples of vertices. This type of clustering is the simplest one can think of. Various other higher order clustering coefficients have been proposed, notably the k-clustering coefficient [20] that takes into account neighbours of distance up to k, coefficients that account for cliques of size larger than 3, cycles and other motifs, see [13]. In our work we will only consider the clustering of triadic closure quantified by the clustering coefficient above because it is much easier to work with analytically and is also the most common.

1.1.3 Random graph models

We will now give a brief overview of the most common random graph models talked about in the networks literature.

Erdős and Rènyi

This is probably the simplest form of a random graph. It is also the most studied and whose structure we know most about [3]. This random graph undertakes a phase transition at the point $p = \frac{1}{n}$. Above this point we have a unique giant component and all other components are small. It is also well known that it has a zero clustering in the limit of large graph size.

Because every edge in the graph is present independently with a probability p. The probability that two vertices that share a common neighbour are connected is p. So, any connected triple is closed with probability p and hence the clustering coefficient is also p. In the case of sparse graphs the probability p is taken as a decreasing function of n of the form $\frac{c}{n}$ in order to obtain an average degree of c and a total number of edges of order $\theta(n)$. So for sparse graphs, the clustering coefficient is zero in the limit of large graph size.

The configuration model

As described before, in this model the degree sequence of the graph is given as a parameter. Actually, the parameter is usually a degree distribution function p_k from which we create a degree sequence d_k giving the degree of each vertex. Given a degree sequence, a random graph is constructed by uniformly selecting a graph among all possible graphs with this degree sequence.

The configuration model has been studied for quite sometime now [22, 23]. Many of its properties are known including the criterion for the formation of a giant component, the number of its cycles and its size. Some of these results were derived rigorously like [9], others using heuristics and approximations. Newman, for example, exploits the tree like structure in such random graphs to derive many properties using the generating function formalism [24], which as we will see later, can also be adapted to the generalised form of the configuration model that we discuss here.

Growing networks

One can classify models of random graphs into two types: *Static* models and *growing* models. In static models the number of vertices is fixed, a graphs is then selected at random from a class of graphs with that size. In growing networks vertices and edges are gradually added to the graph. In static models the aim to mimic or recreate properties of real world graphs, whereas in growing networks the aim is to explain why networks are the way they are, by explaining how they grow.

Some of the most popular models in the category of growing networks are those aimed at explaining the right skewed degree distributions of real world networks described previously. Some of these like Price's model, are actually models of directed graphs but are still worth mentioning.

Price's model [33] is a model that was originally aimed at explaining the power law distribution found in the in-degree of scientific paper citation networks. This model relies on a property called preferential attachment in which newly added vertices are more likely to attach to vertices with high in-degree. This is motivated by the intuition that an already highly cited paper is more likely to be encountered by the author of a new paper and therefore be cited again.

The graph is constructed by adding one vertex at a time with mean out-degree m, and each edge it is connected to a vertex of in-degree k with probability

$$\frac{(k+1)p_k}{\sum_k (k+1)p_k} = \frac{(k+1)p_k}{m+1}$$

where p_k is the in-degree distribution of the graph. This probability is proportional to $(k+1)$ to give a chance to newly created vertices which have in-degree zero. The degree distribution p_k is calculated using a method from statistical physics called the *master equation method* that aims to find a stable point in p_k, in the limit of large graph size, it has been shown in [33] that $p_k \sim k^{2+\frac{1}{m}}$.

Another popular model in this category, is the model by Barabási and Albert [1] that endavours to explain the the degree distribution of pages in the World Wide Web network. Similarly, it uses a linear preferential attachment property, but the graph here is undirected contrary to the network which it tries to model (The World Wide Web is directed). Added vertices have fixed degree m and the attachment property is proportional to the degree of target vertices. The probability that a new vertex is a vertex of degree k is

$$\frac{kp_k}{\sum_k kp_k} = \frac{kp_k}{2m}$$

This model can also be solved using the master equation method and has a degree distribution of $p_k \sim k^{-3}$ in the limit of large k. This result was subsequently derived using more rigorous methods by Bollobás [7].

Clustering models

Since they are main topic of this dissertation, it is also worth mentioning some random graph models that were aimed to create graphs with non zero clustering coefficient in the limit of large graph size.

There are many growing network processes such as the ones described above that involve the addition as well as the deletion and moving around of edges. One category of these models aims to create clustered graphs using *triadic closure processes*[16]. In these models, in addition to preferential attachment of newly added vertices one tries to add edges to form closed triangles. These models however all seem not to be tractable, and the calculation of their properties is limited to numerical methods.

The small world model proposed by Watts and Strogatz [34], is based on the process of rewiring the edges of a regular lattice or ring, this model produces non zero clustering coefficients and a small average distance between vertices, hence the name. Its main criticism however, is that it produces homogeneous degree distributions which is a lot unlike real world graphs.

Finally, two very recent models by Gleeon [14] and Newman [29] have been shown to have non zero clustering and many of their properties have been computed. The authors show that making certain assumptions of the structure of the graph (the tree like structure) makes the calculation of certain properties very simple.

We must also acknowledge the very recent paper by Bollobás, Janson and Riordan [5] in which they present a very general and flexible model that allows clustering and is also tractable. This work is still very recent and not much work has been done on it in terms of applications.

Chapter 2

The classical model

Before we start discussing random graph models in more detail. We need to define certain key concepts that are of particular importance to our work here.

Definition 2. *We say that an event A_n happens with high probability, and we denote it whp , iff*

$$\lim_{n\to\infty} P(A_n) = 1$$

So an event A_n happens whp in the context of a graph G, if it happens with probability 1 in the limit of large graph size.

Definition 3. *The giant component C_1 is the unique component whose fractional size tends to a non zero constant in the limit of large graph size i.e. :*

$$\lim n \to \infty \frac{|C_1|}{n} = c > 0$$

where $|C_1|$ denotes the size of C_1 and n is the size of the graph.

The term giant component was first used in the context of the Erdős and Rènyi model to designate the unique component whose size was $\theta(n)$, this implied that when a giant component appears all the other components are of size $o(n)$. This term was carried on to other models like the configuration model and it designates a component with these same properties.

Definition 4. *A small component C is a small component if it is not giant, that is it contains a zero fraction of the vertices of the graph in the limit of large graph size:*

$$\lim n \to \infty \frac{|C|}{n} = 0$$

Definition 5.

- *The diameter of a graph is the longest geodesic distance between any two vertices in the graph. That is it is the longest shortest path in the graph.*
- *The girth g of a graph is the length of its shortest cycle.*

Definition 6. *We say that a graph, or a component of a graph is locally tree like if it has diameter d and girth g such that $2d \leq g$.*

So roughly speaking, a component is locally tree like if does not have any short cycles. This is illustrated by figure (2.1).

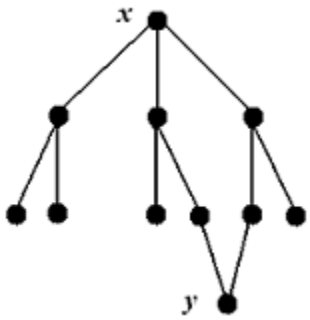

Figure 2.1: Example of a locally tree like graph.

2.1 The configuration Model

Before we move on to describing Newman's model which is the main model on which this dissertation is centered, it is essential to be familiar with the model on which it is based namely the configuration model.

Definition 7. *The degree sequence of a graph G, is a sequence of integers $d_1, d_2, \ldots$ such that d_k is the number of vertices of degree k.*

Definition 8. *An asymptotic degree sequence is a sequence $d_1(n), d_2(n), \ldots$ of integer valued functions such that for a fixed graph size n, we obtain a fixed degree sequence $d_1, d_2, \ldots$.*

In the configuration model we are given an asymptotic degree sequence $d_1(n), d_2(n), \ldots$, which for a given graph size n gives the degree sequence $d_1, d_2, d_3, \ldots$. This creates a space Ω of all possible graphs of size n with such a degree sequence. We construct our random graph by uniformly selecting a member of Ω.

We say that a degree sequence is *feasible* if the set Ω is non empty. We say that an asymptotic degree sequence is *smooth* if there exists a limiting degree distribution p_k such that

$$\lim_{n\to\infty} \frac{d_k(n)}{n} = p_k$$

Definition 9. *We say that an asymptotic degree sequence is well behaved if it is feasible, smooth and*

$$\lim_{n\to\infty} \sum_{k\geq 1} \frac{k(k-2)d_k(n)}{n} = \sum_{k\geq 1} k(k-2)p_k$$

where p_k is the limiting degree distribution.

Definition 10. *We say that an asymptotic degree sequence is sparse if there exists a constant K such that*

$$\sum_{k\geq 0} k\, \frac{d_k(n)}{n} = K + o(1)$$

In their paper Molloy and Reed [22] have derived and proven, that under certain conditions, the criterion for the formation of a giant component is $\sum_{k\geq 1} k(k-2)p_k > 0$. Their result is stated as follows :

Theorem 1. (Molloy & Reed) *Let $d_k(n)$ be a well-behaved sparse asymptotic degree sequence, such that for any $\epsilon > 0$, if $k > n^{1/4-\epsilon}$ then $d_k(n) = 0$. Let G be a graph of size n with degree sequence $d_k(n)$ chosen uniformly at random from the space of all such graphs. Then:*

- *If* $\sum_{k\geq 1} k(k-2)p_k > 0$ *then there exist constants* $c_1, c_2, c_3 > 0$ *such that* G *almost surely has a component with at least* $c_1 n$ *vertices and* $c_2 n$ *cycles. Furthermore, if* $\sum_{k\geq 1} k(k-2)p_k > 0$ *is finite then* G *almost surely has exactly no other component with size greater than* $c_3 \log n$.
- *If* $\sum_{k\geq 1} k(k-2)p_k < 0$ *and there exists an* $\epsilon > 0$ *and a function* $w(n)$ *such that* $0 \leq w(n) \leq n^{1/8-\epsilon}$ *and if* $k \geq w(n)$ *then* $d_k(n) = 0$ *for all* n. *Then there exists a constant* R *such that* G *almost surely has no component with size greater than* $Rw(n)^2 \log n$ *vertices and more than one cycle.*

This result shows that small components have a tree like structure. The condition required on the maximum degree to be at most $n^{1/4-\epsilon}$ is required here simply to guarantee that the algorithm used to construct a random configuration produces a simple graph with positive probability. The remainder of the results are therefore conditioned on the graph being simple. More recently, Janson [18] has shown that no constraint on the maximum degree is required to satisfy this, but simply that the second moment of the degree sequence increases linearly in n. The proof of theorem 1 is based on the following algorithm used to construct a random configuration for a given fixed degree sequence d_k.

Algorithm 1. *1. Create a set* S *such that for each vertex* i *with degree* k, S *contains* k *copies of* i. *We create a random configuration* F *by pairing up the copies in* S. *If a copy of a vertex* i *is added to* F *we say that* i *is* ***exposed*** *and the remaining copies corresponding to* i *that have not been added to* F *are said to be* ***open****.*

2. Repeat until S *is empty :*

- *Pick an element from* S *selected uniformly at random, choose its partner similarly at random, add the pair to* F *and remove it from* S.
- *Repeat until there are no open copies left*
  - *Choose an open copy from* S *and pair it with any element from* S, *add the pair to* F *and remove them from* S.

In other literature, the same configuration building process is described in terms of pairing up stubs or half edges [3]. Note that step 1 is only executed when starting a new component and that as long as as there are open vertex copies we are still exposing the same component.

The proof of the result itself is based on the intuition that the initial rate of increase of the number of open vertex copies is roughly $\sum_{k\geq 1} k(k-2)p_k$. If this is positive then we expose a large number of vertices in our component. If it is negative the number of open vertices goes to zero very quickly and we don't expose many vertices.

Further properties have been proven for the configuration model. In a subsequent paper [23], the same authors calculated and proved that the size of the giant component C_1 is $\epsilon n + o(n)$ for some some ϵ dependant on the degree sequence. They also determine $\lambda_1, \lambda_2, \ldots$ such the structure of the graph after removing C_1 is that of another random graph of size $n' = n - \epsilon n + o(n)$ and degree sequence such that $\lambda_i n'$ vertices have degree i.

In a more recent paper[18], Janson and Luczak have produced a new proof of the criterion of the emergence of the giant component in thoerem (1), and that all other components are small. They used a different method from Molloy and Reed that relies on the convergence of random variables. Most importantly, they do not require a limit on the maximum degree of the order $n^{-1/4}$, but a simple condition on the second moment of the asymptotic degree distribution: $\sum_i d_i^2 = O(n)$. However, in their paper they did not produce any results about the number of cycles or size of small components.

The configuration model was generalised to bipartite graphs by Neman et al [26]. It was also shown by Dorogovtsev et al that almost all vertices of the configuration model are mutually equidistant [11].

2.2 Newman's generating functions

Since the work done by Molloy and Reed, many further properties of the configuration model have been calculated. These results build on those shown by Molloy and Reed in the sense that they use heuristics and approximations that assume the tree-like structure proved by them.

One of these methods is the probability generating function formalism developed by Newman [24]. This method exploits the tree-like structure of the graph, that is the property that the graph has very few cycles in the limit of large graph size, allowing the neighbours of a given vertex to be independent. This allows the computation of key properties through the iteration of generating functions which is justified by the power property.

The power property of generating function

If $X_1, \ldots, X_n$ is a sequence of independent random variables (not necessarily from the same distribution) and

$$S_n = \sum_{i=1}^{n} a_i X_i$$

Then the probability generating function of S_n is given by

$$G_{S_n}(x) = E(x^{S_n}) = E(x^{a_1 X_1 + \cdots + a_n X_n}) = E(x^{a_1 X_1} \ldots x^{a_n X_n})$$

By independence this is

$$= E(x^{a_1 X_1}) \ldots E(x^{a_n X_n}) = G_{X_1} \ldots G_{X_n}$$

Similarly if $X_1, \ldots, X_N$ is a sequence of independent random variables, but identically distributed with generating function $G_X(x)$, where N is an independent random variable itself. Then

$$G_{S_N}(x) = E(x^{S_N}) = E[E(x^{S_N})|N] = E[\big(G_X(x)\big)^N|N] = G_N\big(G_X(x)\big)$$

We give a short illustration taken from Newman's paper [24] to show how generating functions are used to compute key properties in the classical configuration model.

Suppose for instance that we are given the degree distribution function p_k. We would like to compute the average number of vertices which are a distance two away from a random vertex v in a random graph with this degree distribution. Given that the graph has very few short cycles, the picture around v looks approximately like figure (2.2).

Essentially, the number of second neighbours of v is the sum of the number of neighbours of its immediate neighbours. The number of immediate neighbours is generated by the function

$$g_p(x) = \sum_k x^k p_k \tag{2.2.1}$$

The number of the neighbours of a neighbour of v is however not generated by the same function. It is intuitive to see that if we select a random edge in the graph and follow it in either direction, the probability that we land on a vertex of degree k is proportional to its degree:

$$Pr(\text{one of its incident edges is selected})\ p_k = \frac{k}{\sum_k k p_k} p_k$$

Note also that selecting a random vertex then following one of its edge at random is equivalent to selecting a random edge in the graph. What we are interested in here, is the probability that

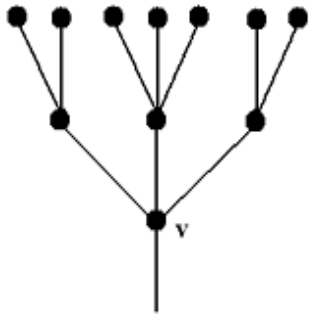

Figure 2.2: No short cycles around a randomly selected vertex.

a given vertex reached by selecting a random edge and following it has degree k not counting the edge that we just traversed. This happens with probability

$$q_k = \frac{(k+1)}{\sum_k k p_k} p_{k+1} \tag{2.2.2}$$

This is referred to as the *excess degree*. Let X be the random variable representing the number of neighbours of v as in fiigure (2.2). Let $Y_1, Y_2, \ldots, Y_X$ be the sequence of random variables for the number of neighbours of each of the neighbours of v, not counting v.

Then the total number of second neighbours is $Y_1 + Y_2 + \cdots + Y_X$. Therefore its generating function is $G_X(G_Y)$. Since the Y_is are independent because their vertices are disjoint and X is independent of the the Y_is because we do not count the edges between v and its neighbours. Since X has distribution p_k and the Y_is have distribution q_k, the number of second neighbours is generated by $g_p(g_q(x))$. Note also that:

$$g_q(x) = \frac{1}{\sum_k k p_k} \sum_k (k+1) p_{k+1} x^k = \frac{1}{\langle k \rangle} g_p'(x) \tag{2.2.3}$$

where $\langle k \rangle$ denotes the average degree of vertices. We find that the expected number of second edges is given by

$$g_p(g_q(1))' = g_p'(1) g_q'(1) = \langle k \rangle \frac{1}{\langle k \rangle} g_p''(1) = g_p''(1)$$

For example if we had a Poisson degree distribution $Po(c)$. The generating function would be $g_p(x) = e^{c(x-1)}$, so the average number of second neighbours is $g_p''(1) = c^2$.

Many more examples of other properties computed using the generating function formalism are given in [24]. It is important to note here that what is most impressive about Newman's generating functions formalism is that it allows the derivation of key properties not just for a random graph with one specific degree distribution but for any degree distribution given that we can solve equations (2.2.1), (2.2.3) about its generating function, which if we can't do analytically we can achieve numerically.

Remark 1. *It is important to note here that the generating function formalism requires that the first neighbours are independent and therefore disjoint. This is why we require the condition that components are locally tree like. Where components are not completely tree like this method provides only an approximation and not an exact solution.*

Remark 2. *It is also important to note that the Molloy and Reed result shown in thoerem (1) only shows that small components are tree like and not the giant component. In fact it has many cycles. In his papers [29, 26], Newman nonetheless uses this method to estimate properties about the giant component. This is justified by the subsequent work by Molloy and Reed [23] on the size of the giant component which we shall not discuss in this dissertation.*

Finally, to motivate the next section we will use the generating functions method to show the weakness of the of the configuration model that the new model by Newman presented in the following chapter tries to fix, namely that a random graph with a fixed degree distribution has zero clustering in the limit of large graph size [30].

Consider a randomly selected vertex v, we will attempt to estimate the clustering coefficient by estimating the average probability that two pair of its neighbours are connected. This is illustrated in figure (2.3).

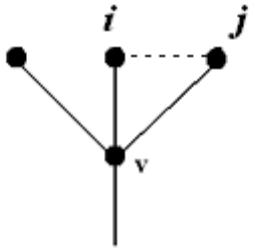

Figure 2.3: Clustering in the configuration model.

Suppose v has two neighbours i and j with excess degrees k_i, k_j, the probability that they are connected given that the graph is constructed by pairing half edges at random is:

$$\frac{k_i k_j}{n\sum_j k p_k} = \frac{k_i k_j}{n\langle k\rangle}.$$

because the total number of half edges is $\sum_j k(np_k)$. Therefore the mean probability that i and j are connected is:

$$\frac{\langle k_i k_j\rangle}{n\langle k\rangle} = \frac{\sum_{i,j} q_i q_j k_i k_j}{\langle k\rangle} = \frac{(\sum_i q_i k_i)^2}{n\langle k\rangle}$$

The mean excess degree $\sum_i q_i k_j$ is given by:

$$\frac{\sum_i p_i k_i(k_i-1)}{n\langle k\rangle} = \frac{\langle k^2\rangle - \langle k^2\rangle}{n\langle k\rangle}$$

So the mean probability that that i and j are connected is:

$$\frac{(\langle k^2\rangle - \langle k^2\rangle)^2}{n\langle k\rangle^3} = \frac{\langle k\rangle}{n}\frac{(\langle k^2\rangle - \langle k^2\rangle)^2}{\langle k\rangle^4}$$

For sparse graphs, the value $\langle k\rangle$ is constant and $\langle k^2\rangle \leq \langle k\rangle^2$. Therefore, the above fraction tends to zero as n goes to infinity and therefore the clustering coefficient is zero in the limit of large graphs size for a random graph with fixed sparse degree distribution.

Chapter 3

Newman's random graph model with clustering

3.1 Newman's random graph model with clustering

In a very recent paper [29], Newman introduced a random graph model which has a provable non zero clustering coefficient in the limit of large graph size. This model can be considered a generalisation of the classical configuration model, in the sense that the classical configuration model is a special case of the new one.

In Newman's model we specify two kinds of vertex specific sequences: one sequence for the number of single edges incident to a vertex i denoted s_i, and one for the number of triangles in which the vertex participates in denoted t_i. In this way a vertex has total degree $s_i + 2t_i$. The motivation behind this model is that for a significant number of triangles attached per vertex, the fraction of closed triples will tend to a non zero value.

An example graph created in this way is shown in figure (3.1). The shaded triangles are those explicitly specified by the degree sequence.

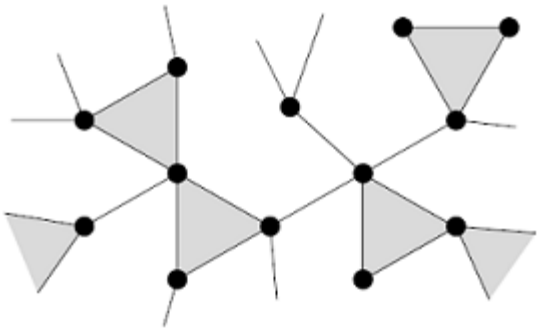

Figure 3.1: Newman's random graph model with clustering. Source [29].

In this model we define a joint degree distribution function p_{st} which represents the probability that a randomly chosen vertex has s single edges attached to it and t triangles. Note that if we were to define two separate degree distributions p_s and p_t we would loose any correlation that we would like to implement between the number of triangles and the number of singles edges a vertex can have.

In his paper, Newman derives many key properties for this model using the same generating

function formalism that he uses with the classical configuration model with the only difference that we now have a joint degree distribution. In doing this, he is implicitly assuming the same locally tree like structure of the classical model i.e. that the graph contains very few short cycles.

Of course, we do have explicitly implemented triangles which form short cycles. This implies that by using the generating functions he is assuming that the graph is locally tree like in the sense that it does not contain short cycles not counting those explicitly specified by the degree sequence. In the next section we will define clearly what we mean by short cycles excluding explicit triangles.

However, regardless of what these short cycles are, it is not necessarily clear that we still won't have any short cycles even if it has been proved for the classical configuration model in the limit of large graph size. A skeptic can think that with triangles explicitly attached to vetices we are more as likely to form cycles.

Furthermore, Molloy and Reed's work does not just give a qualitative description of the behavior of the graph. As mentioned earlier in theorem (1), in their work, they give a clear criterion for the point where the giant component forms. They also give bounds for the size of smaller components and the number of cycles within components including the giant component. It would be of interrest to have equivalent results for Newman's model. This will be the main result of the next chapter.

Chapter 4

A proof for the new model

The aim of this chapter was to give a proof for the formation of a giant component and the locally tree like structure of the small components of a random graph with a fixed joint degree sequence consisting of single edges and triangles as in the model introduced by Newman. However, Having found some important mistakes in the proof at a very late stages of this dissertation and because of the fact that the main result (the criterion for the formation of a giant component) does not agree with Newman's predictions using the generating functions method, we decided to omit this proof here and put in the appendix with some notes on what mistakes were made.

Instead, we dedicate this chapter to provide a proof of a special case of the above random graph model, which is a random graph with no single edges i.e. a random graph with a fixed triangle sequence.

Although not a very straighforward one, this proof is essentially an adapted from of the classical configuration model presented by Molloy and Reed. We will in fact follow the same structure and order in which it is presented to facilitate it to a reader familiar with Molloy and Reed's paper [22]. We have also made some corrections to their proof, there given in the form of lemmas (4) and (8).

We aim by this chapter to show how Molloy and Reed's proof of the classical configuration model can be adapted to other types of degree sequence, in this case triangles, and possibly motivate the reader to research the problem that we attempted originally, which is to prove that random graph with a fixed joint distribution and triangles behaves similarly to the classical fixed degree distribution configuration model in that it has a threshold below which all components are small and tree like and above which a giant component forms and contains many cycles. Most importantly, one need to define (and prove) a criterion or the transition point where this happens.

We will start by defining a few concepts, we will then state our result in the form of a theorem. We will split the theorem into smaller lemmas which we will then prove individually.

Definition 11. *A triangle degree sequence is a sequence of non negative integers $t_1 \dots t_n$ that represents the number of triangles t attached to the vertex i. We will usually denote this by (t_i).*

Since we are dealing only with triangle degree, we will simply refer to them as degree sequences.

Definition 12. *An asymptotic triangle degree sequence is a sequence of integer valued functions $t_1(n), t_2(n) \dots$, such that for a fixed graph size n we obtain a fixed triangle degree sequence (t_i).*

Definition 13. *We say that a triangle degree sequence is feasible if the set of all possible graphs with that sequence is non-empty.*

4.1 A configuration model with triangles

Suppose we are given an asymptotic triangle degree sequence $t_1(n), t_2(n), \ldots$ representing the number of triangles for a each vertex i in a graph of n vertices.

Using this sequence, we construct the sequence $d_j(n)$, where each entry represents the number of vertices in the graph with exactly j triangles.

First, we Construct a set S of copies of vertices for our graph, by creating t copies for every vertex with t triangles attached to it. In total we have $\sum_i t_i(n) = \sum_{j\geq 1} d_j(n)$ copies of vertices in S.

Definition 14. *A random configuration F is a partition of S that consists of a set of triples of the copies of S. This partition is selected uniformly at random from the set of all possible partitions.*

Definition 15. *An configuration cycle is a cycle that is not created explicitly by a triangle in the degree sequence.*

We will construct a random graph G with the above degree sequence by constructing a *random configuration* using the following algorithm.

Algorithm 2. *We construct our configuration F by tripling the copies of the set S. We say that a vertex is exposed if any of its copies has been added to F, and we say that the copies of an exposed vertex that remain in S are open.*

Repeat the following until S is empty:

1. *Expose a random vertex v in G by selecting a random element or copy in S then exposing all the remaining copies of the same vertex.*
2. *Repeat the following until there are no open copies left.*
   - *Select an open copy x from S uniformly at random. Then, select another copy y from S (open or non open) uniformly at random. If y is not an open copy, expose its corresponding vertex by opening all its remaining copies. Remove y and half of x from S.*
   - *Select a copy z from S uniformly at random. Remove z and the other half of x from S and add the triple (x, y, z) to F. If z is a copy of an unexposed vertex, open all its remaining copies.*

We can see that using this algorithm, we construct any configuration with the specified degree sequence uniformly at random from the set of all possibilities. The action of tripling the three copies (x, y, z) corresponds to connecting three vertices in a triangle. Hence, the algorithm is exposing the components of G one at a time. A component is fully exposed when there are no more open copies left and a new component is started everytime we go back to step 1. Note also that in step 1, the vertex did not have to be selected at random, it could be any vertex whose component we would like to expose.

Of course, the above algorithm essentially constructs a multi-graph, but for the case of the classical configuration model Janson [17] showed that given certain conditions, there is a positive probability that the graph is simple. This condition is that the second moment of the degree sequence is at most linear in the size of the graph n :

$$\sum_i s_i(n)^2 = O(n).$$

Where $s_i(n)$ represents the degree of the ith vertex. This is an improvement to the condition used by Molloy and Reed which is a restriction that the maximum degree of the graph is $n^{-1/4-\epsilon}$ for any $\epsilon > 0$.

We claim that one can show the same result under similar conditions for a random graph with triangle degree sequence and larger motifs beyond triangles. Although we do not show this here, in what follows we condition on the fact that graph we obtain is simple. We will state all our lemmas in terms of results for random configurations this will imply that the results hold for a random graph.

Definition 16. *We say that a joint degree sequence is sparse if the sum of all degrees of the vertices of the graph is linear in the size of the graph:*

$$\sum_{i\geq 1} t_i = \sum_{j\geq 1} j d_j = Kn + o(n).$$

Definition 17. *We say that a triangle degree sequence is well-behaved if it is feasible and there exists constants p_j such that*

1. $$\lim_{n\to\infty} \frac{d_j(n)}{n} = p_j.$$

2. *For all j, $j(2j-3)d_j(n)/n$ tends uniformly to $j(2j-3)p_j$ as $n \to \infty$.*

3. *$\lim_{n\to\infty} \sum_j (j)(2j-3)\frac{d_j(n)}{n}$ tends uniformly to $\lim_{n\to\infty} \sum_j (j)(2j-3)p_j$, i.e. for all $\epsilon > 0$, there exists j^*, N such that for all $n > N$:*

$$|\sum_{j}^{j^*} (j)(2j-3)\frac{d_j(n)}{n} - \sum_j j(2j-3)p_j| < \epsilon$$

Definition 18. *We define, the following constants*

$$D = \sum_j j(2j-3)p_j \tag{4.1.1}$$

$$Q = \frac{\sum_j j d_j (2j-3)}{\sum_j j d_j} = \frac{D}{K} \tag{4.1.2}$$

The expression D represents the criterion for the formation of the giant component. Note that this value can be infinite. We now state our main result.

Theorem 2. *Let $t_1(n), t_2(n), \ldots$ be a sparse, well-behaved asymptotic triangle degree sequence such that the probability that a random configuration with this sequence constructs a simple graph is positive. Let G be a graph with the above triangle degree sequence chosen uniformly at random from the set of all graphs with such a sequence. Then*

1. *If $D > 0$ and if Q is finite. Then, there exist constants $c_1, c_2, c_3 > 0$ such that G whp has one component with at least $c_1 n$ vertices and $c_2 n$ configuration cycles. Furthermore, G whp has exactly no other component with size greater than $c_3 \log(n)$ and no such component has more than one configuration cycle.*

2. *If $D < 0$ and there exists an $\epsilon > 0$ and a function $w(n)$ such that if $0 \leq w(n) \leq n^{1/8-\epsilon}$ and the maximum degree of the sequence is at most $w(n)$ for all n. Then there exists a constant R such that G with high probability has no component with size greater than $Rw(n)^2 \log n$ vertices and such component has more than one configuration cycle.*

Remark 3. *Note that the second condition required in the first case of the theorem i.e. that Q is finite can be made redundant if one is able to show, as Janson did [18] for the classical configuration model, that the condition required a random configuration forms a simple graph with positive probability is that the second moment of the degree is $O(n)$. This would give :*

$$Q = \frac{\sum_j jd_j(2j-3)}{\sum_j jd_j} \leq \frac{Ln}{Kn} = \frac{L}{K}$$

Note that Q represents the initial rate of increase of open vertices as we begin our exploration of any component. The sign of Q is determined by D. We motivate the main idea behind the proof as follows: If the initial rate of increase of open copies is positive, then we are likely to expose many vertices and form a giant component. If it is negative the number of open copies goes quickly to zero and we expose a small component. Let us first we define few variables that will be useful later.

Definition 19. *We define the following variables :*

- *Let X_r be the number of open copies after the r^{th} pair has been formed. Note that a triple here is counted as two pairings. When we say that a pair has been exposed we mean an execution of one of the two sub steps of step 2 of algorithm (2).*
- *Let C_r be the number of components fully or partially exposed when the r^{th} pair has been exposed, again counting a triple as two pairs.*
- *We say that a back-edge has been formed when we pair an open copy of S with another open copy of S in step 2. This in fact corresponds to forming a configuration cycle.*
- *Let Y_r be the number of back-edges formed when the rth pair has been exposed.*

We will now motivate the remainder of our proof by looking at the initial rate of increase of open copies X_r. This is given by:

$$\sum_j \frac{jd_j}{\sum_j d_j}(j - \frac{3}{2}).$$

This is because every copy in S is selected with a probability $(j/\sum_j jd_j)$, and by doing so we add j new open copies and remove $\frac{3}{2}$.

Remark 4. *Note that it would have been more intuitive to define our construction algorithm (2) by selecting two open copies y, z in one step, tripling them with the open copy x then removing all three copies and exposing the two new vertices. Both constructions are in fact equivalent. If we construct our configuration in way the initial rate of increase of open vertices would have been:*

$$\sum_{i,j} \frac{id_i}{\sum_i id_i} \frac{jd_j}{\sum_j jd_j}(i + j - 3)$$

because we expose two new vertices with degrees i, j respectively. This sum is:

$$\begin{aligned}
&= \frac{1}{\sum_i id_i \sum_j jd_j} \sum_{i,j} id_i jd_j (i + j - 3) \\
&= \frac{1}{\sum_i id_i \sum_j jd_j} [2 \sum_{i,j} (i^2 d_i jd_j) - 3 \sum i, jid_i jd_j] \\
&= \frac{1}{\sum_i id_i \sum_j jd_j} [2 \sum_i (i^2 d_i) \sum_j (jd_j) - 3 \sum_i (id_i) \sum_j (jd_j)] \\
&= \frac{1}{\sum_i id_i \sum_j jd_j} \sum_j (jd_j) [2 \sum_i (i^2 d_i) - 3 \sum_i (id_i) \sum_j (jd_j)] \\
&= \frac{1}{\sum_j jd_j} \sum_i d_i i(2i - 3)
\end{aligned}$$

Which is double the expected increase we have in each sub step of step 2 of algorithm (2), but most importantly it produces the same criterion of the formation of the giant component in theorem (2).

Definition 20. *We define the following useful variables :*

- *We define the variable Z_q to be the sum of $(j - \frac{3}{2})$ over the first q exposed vertices.*
- *We also define the analogous variable W_r to be the sum over $(j - \frac{3}{2})$ over all vertices exposed by the time the rth pair has been exposed, counting a triple as two pairs.*

Remark 5. *The reason we introduce Z_q is that it has the same rate of increase as X_r but behaves much more nicely in that it only increases by $(j - \frac{3}{2})$ every time a vertex with j triangles is exposed. Hence, it is easy to put a bound on it's expected value when a fixed number of vertices have been exposed as we shall see later.*

We now relate all the variables defined previously. We define the variable R_q to be the number of pairs exposed by the time we expose the qth vertex i.e. $W_{R_q} = Z_q$.

Remark 6. *Note that X_r is (roughly) the same as W_r except when we form a back edge. In which case X_r decreases. Hence we obtain :*

$$W_r = X_r + \frac{3}{2} Y_r. \tag{4.1.3}$$

Remark 7. *We can also relate W_r to Z_q. If no back edges are formed we would have exposed $R_q + 1 = q$ or $R_r = r - 1$ vertices. Consequently we would get $W_r = Z_{r-1}$, but given that we get some back edges we have that the number of steps $R_r = r + Y_r - 1$ or $r = R_r - Y_r + 1$, so*

$$W_r = Z_{(r - Y_r + 1)}. \tag{4.1.4}$$

Remark 8. *Z_r decrease by at most $1 - \frac{3}{2} = -\frac{1}{2}$) every time a vertex is exposed. This happens when we expose a vertex with degree 1. Therefore*

$$\begin{aligned}
Z_r &\geq Z_{(r - Y_r + 1)} - \frac{1}{2}(Y_r - 1) \\
&= W_r - \frac{1}{2} Y_r \\
&= X_r Y_r \\
&\geq X_r.
\end{aligned}$$

So Z_r is bounded below by X_r.

4.2 Small components

We now show that if the conditions of the second case of theorem (2) are satisfied, the graph has no components of size larger than $\alpha = Sw(n)^2 log(n)$ vertices. We will show that if the expected increase in Z_q is negative, then the probability that it remains greater than zero for too long is very small and therefore the probability that the probability that the number of open copies X_r is greater than zero is also very small.

Lemma 1. *Let F be a configuration that satisfies the conditions of the second case of the theorem. Let v be any vertex then the probability that v lies in a component of size $\alpha = Sw(n)^2 log(n)$ is less than n^{-2}.*

Proof. Suppose that we start our algorithm by exposing v at step 1. We have that

$$Q = \frac{D}{K} = \sum_{i,j} \frac{d_j j}{\sum_j j d_j}(j - \frac{3}{2}) < 0.$$

The probability that a given component has size at least α is at most the probability that $X_\alpha > 0$, which is consequently at most the probability that $Z_r > 0$ from remark (8). This is because if $X_r = 0$ then we would have exposed the whole component.

Initially the rate of increase of Z_r is

$$\sum_j \frac{d_j j}{\sum_j j d_j}(j - \frac{3}{2}).$$

After exposing $q \leq \alpha$ vertices, the rate of growth of Z is highest if the first q vertices that were exposed have degree 1. This because the negative terms in sum $(j - \frac{3}{2})$ are those where $j = 1$. Hence the rate of increase of Z_q is at most:

$$\frac{-\frac{1}{2}(d_1 - q) + \sum_{j\geq 2} d_j j(j - \frac{3}{2})}{(d_1 - q) + \sum_{j\geq 2} j d_j}, \tag{4.2.1}$$

this is at most:

$$\leq \frac{\sum_j d_j j(j - \frac{3}{2})}{\sum_j j d_j \ - q} + \frac{\frac{1}{2}q}{\sum_j d_j j - q}$$

$$\leq \frac{\sum_j d_j j(j - \frac{3}{2})}{\sum_j d_j j} + \frac{\frac{1}{2}q}{\sum_j d_j (i + j) - q}.$$

Because $q \leq \alpha = o(n)$ and $\sum_j d_j(i + j) = Kn = \theta(n)$, we get the expected increase after q steps is:

$$\leq Q + o(1) \leq \frac{3Q}{4} < 0$$

for n large enough. The expected increase in Z_q is still negative, indicating that the process should die out quickly. Given that the degree of the the first chosen vertex v is at most $w(n)$, we get that after α vertices the expected value of Z_α is at most

$$\text{Initial value} + (\text{Rate } \times \alpha) \leq (3Q/4)\alpha + w(n).$$

Because $\alpha = Sw(n)^2 log(n)$, for n large enough it follows that

$$\frac{3Q}{4}\alpha + w(n) \leq \frac{Q}{2}\alpha.$$

We now introduce an important result known as Azuma's inequality that will help bound the probability of Z deviating too far from its mean.

Azuma's inequality

Let $X_0, \ldots, X_n$ be a martingale with $|X_i - X_{i-1}| \leq 1$, for all $0 \leq i < n$, with Let $\lambda > 0$ it follows that

$$Pr(|X_n| > \lambda\sqrt{n}) < e^{-\lambda^2/2}.$$

Azuma's inequality yields the following standard corollary.

Corollary 1. *Let $\Sigma = \Sigma_1, \ldots, \Sigma_n$ be a sequence of random events. Let $f(\Sigma) = f(\Sigma_1, \Sigma_2, \ldots, \Sigma_n)$ be a random variable defined over these events. Then if $E(f|\Sigma_1, \Sigma_2, \ldots, \Sigma_i)$ is $c-$Lipshtiz, that is if there exists constants c_i and $c = (c_1, \ldots, c_n)$ such that for all i :*

$$max|E(f(\Sigma)|\Sigma_1, \ldots, \Sigma_{i+1}) - E(f(\Sigma)|\Sigma_1, \ldots, \Sigma_i)| \leq c_i$$

Then

$$Pr(|f - E(f)| > t) \leq 2\exp\left(\frac{-t^2}{2\sum_i c_i^2}\right).$$

□

We will make use of Azuma's inequality by defining Σ_i to indicate the ith vertex to be exposed, for $i = 1, \ldots, \alpha$ and $f(\Sigma) = Z_\alpha$. We also define $E_{i+1}(x) = E(Z_\alpha|\Sigma_1, \ldots, \Sigma_{i+1})$, where Σ_{i+1} is the event that the $(i+1)th$ vertex is x. We would like to bound

$$|E(f(\Sigma)|\Sigma_1, \ldots, \Sigma_{i+1}) - E(f(\Sigma)|\Sigma_1, \ldots, \Sigma_i)|$$

We will do this by first bounding $|E_{i+1}(x) - E_{i+1}(y)|$ for any x, y. Let u, v be any two vertices. Suppose that we are choosing the $(i+1)$st vertex. We are therefore left with $n-i$ vertices. Note that by ignoring u, v, the distribution of the order in which the remaining vertices are exposed is unaffected by the positions of u and v.

Let Ω be the set of the first $\alpha - i - 3$ vertices in this order. Then:

$$Z_\alpha = Z_i + \sum_{\Omega}(j - \frac{2}{3}) + deg(y_1) - (\frac{2}{3}) + deg(y_2) - (\frac{2}{3}).$$

where y_1 is either u or v and y_2 is either u, v or the next vertex in the order. Hence we see that the choice between u and v can only change Z_α by an amount equal to the maximum degree which is $w(n)$. Hence:

$$\max_{x,y}|E_{i+1}(x) - E_{i+1}(y)| \leq w(n).$$

Given the fact that

$$E(f(\Sigma)|\Sigma_1, \ldots, \Sigma_i) = \sum_x Pr(\text{x is chosen})\ E_{i+1}(x) \leq \max_x\ E_{i+1}(x),$$

we get that

$$\begin{aligned} &|E(f(\Sigma)|\Sigma_1, \ldots, \Sigma_{i+1}) - E(f(\Sigma)|\Sigma_1, \ldots, \Sigma_i)| \\ &\leq \max_{xy}\ |E_{i+1}(x) - E_{i+1}(y)| \leq w(n). \end{aligned}$$

Therefore, the probability that $Z_\alpha > 0$ is at most

$$Pr(|Z_\alpha - E(Z_\alpha)|) > E(Z_\alpha).$$

By Azuma's inequality, this is at most

$$2\exp(-\frac{(Q/2\alpha)^2}{2\sum_i w(n)^2}) = 2\exp(-\frac{(Q/2\alpha)^2}{2\alpha w(n)^2})$$

Substituting $\alpha = Sw(n)^2 log(n)$, we get

$$2\exp\left(-\frac{(Q^2/4Slog(n))}{2}\right) = 2n^{-Q^2/8S} < n^{-2}.$$

The last inequality holds by substituting $S = \frac{17}{Q^2}$ and working through.

Hence the probability that a randomly chosen vertex lies on a component of size at least α is $o(n^{-1})$ and hence the expected number of such vertices is $o(1)$, so with high probability all components in F have size at most $Sw(n)^2 \log(n)$.

4.2.1 Very few configuration cycles

We we will now show that there is asymptotically no component with more than one configuration cycle, when the conditions of the second case of the theorem are satisfied. We will do this by showing that asymptotically we have very few back-edges.

We will build on the result of the last section. We will show that if the size of any component is at most $\alpha = Sw(n)^2 \log(n)$ then the probability that we form two back edges before exposing more than α vertices is very small.

Remark 9. *Looking at our algorithm, we see that X_r the number of open vertices decreases by at most $\frac{3}{2} < 3$ at every execution of step 2. By lemma (1), the size of any component is at most α. This implies that $X_r < 3\alpha$ at any step r during the execution of our algorithm. More precisely, at any step $r \leq \alpha$ we must have that $X_r < 3(\alpha - r)$.*

Lemma 2. *Let F be a configuration satisfying the conditions of the second case of the theorem. Then whp F has no components with 2 cycles.*

Proof. Fix any vertex v. We start our algorithm at step 1 with this vertex. Remember that this is legitimate. We suppose that v lies in a component with more than one cycle. We will show that this happens with a very small probability and therefore asymptotically no such vertices are expected to exist.

Because at every iteration of algorithm (2) we either expose a new vertex or form a back edge, we must have that the second back edge is formed before $(\alpha + 2)$ steps or else we would have exposed more than α vertices which we saw by lemma (1) has a probability less than n^{-2} of happening.

Suppose the first and second back edges are formed at step A and B respectively such that $0 \leq A \leq B \leq \alpha + 2$. The probability that we form two backedges is at most

$$\sum_{A=0}^{\alpha+1} \sum_{B=A}^{\alpha+2} \left(\frac{3(\alpha - A)}{Kn - 3(\alpha - A)}\right)\left(\frac{3(\alpha - B)}{Kn - 3(\alpha - B)}\right).$$

This is because the number of open vertices is less than $3(\alpha - A)$ or $3(\alpha - B)$ and the number of elements left in S is more than $Kn - 3(\alpha - A)$ or $Kn - 3(\alpha - B)$. This probability is at most:

$$\sum_{A=0}^{\alpha+1}\sum_{B=A}^{\alpha+2}\left(\frac{3(\alpha-B)}{Kn-3(\alpha-B)}\right)^2$$
$$\leq \sum_{A=0}^{\alpha+2}\sum_{B=0}^{\alpha+2}\left(\frac{3(\alpha-B)}{Kn-3(\alpha-B)}\right)^2$$
$$\leq (\alpha+2)\sum_{B=0}^{\alpha+2}\left(\frac{3(\alpha-B)}{Kn-3(\alpha-B)}\right)^2$$
$$\leq (\alpha+2)\frac{1}{(Kn-3(\alpha+2))^2}\sum_{B=0}^{\alpha+2}(3(\alpha-B))^2$$
$$= (\alpha+2)\frac{1}{(Kn-3(\alpha+2))^2}\sum_{B=0}^{\alpha+2}(3B)^2$$
$$\leq (\alpha+2)\frac{9}{(Kn-2(\alpha+2))^2}(\alpha+2)^3.$$

Because $\alpha = Sw(n)^2log(n)$, if we take $w(n) = n^{1/8-\epsilon}$ for any $\epsilon > 0$, we get

$$\frac{9}{(Kn-3(\alpha+2))^2}(\alpha+2)^4 = o(n^{-1})$$

So the probability that a random vertex v lies in a component with two cycles is at most $o(n^{-1})$. Therefore the expected number of these is $o(1)$. So with high probability there are no components with more than once cycle as n tends to infinity.

□

Remark 10. *Note that this way of bounding the probability of forming two back edges is different from the way it is presented in Molloy and Reed's paper [22]. This was done in an attempt to eliminate or at least relax the constraint that we have on the maximum degree in the second case of theorem (2). Because this probability largely depends on α, this was then reduced to giving a better bound for the maximum size of small components. However, we were unable to find a better bound than that given by Azuma's inequality in the Molloy and Reed paper.*

4.3 A giant component

In this section, we will consider the first case of theorem (2). We will show that given that $D, Q > 0$ and finite, then with high probability our graph has a giant component with at least a linear number of cycles.

We will proceed as follows: First, we start our configuration building algorithm with any given vertex, and we show that after a certain number of steps, Z_q is very large with high probability. We will then use relation (4.1.3) to show that X_r is also very large with high probability. Having shown that the number of open vertices X_r is very large, we will deduce that with high probability our configuration building algorithm will form a large number of back edges and exposes a large number of new vertices, hence forming a giant component and a large number of cycles.

Lemma 3. *Let F be a configuration that satisfies the conditions of the first case of theorem (2), then there exists $0 < \epsilon < 1$ and $0 < \Delta < min(\frac{1}{2}, \frac{K}{2})$ such that for all $0 < \delta < \Delta$, then a.s. $Z_{\delta n} > \epsilon\delta n$. Moreover, The probability of the converse is z_1^n for some $0 < z_1 < 1$.*

Proof. In what follows, we will assume for simplicity that δn is an integer. This can be achieved for n large enough.

Recall that in the proof of lemma (1), we bounded the expected increase in Z_q after a number of steps of $\alpha = o(n)$. We then used this to have a bound on the expected value of Z_q itself. Then we showed using Azuma's inequality that with high probability we cannot deviate too far from the mean.

We will proceed similarly. However, the problem here is that we want to bound the expected increase in Z_q after a linear number of steps δn. This causes the probability of choosing a copy of a vertex of a certain degree to shift significantly and therefore the expected increase (4.2.1) in Z_q shifts significantly as well.

To get around this, we will define a new variable Z_q^* that behaves much more nicely than Z_q such that Z_q majorises Z_q^*, i.e. that:

$$Pr(Z_q \geq x) \geq Pr(Z_q^* \geq x).$$

We will then show that Z_q^* grows as largely as we want it. Define q_j to be the initial probability that that we choose a copy of a vertex of degree j. We have that:

$$q_j = \frac{jd_{i,j}}{\sum_j jd_{i,j}} = j\frac{p_j}{K}.$$

This probability is likely to significantly shift after δn steps. We define Z_q^* by fixing a number j^* and a sequence of probability values $\phi_1, \phi_2, \ldots, \phi_{j^*}$. Such that Z_q^* is the sum of all $(j - \frac{3}{2})$ by the time the qth vertex is exposed, with the difference that every vertex of degree j is chosen with the fixed probability ϕ_j at every step, and that if we select a vertex of triangle degree greater than j^* we treat as having triangle degree 1 i.e. subtract $\frac{3}{2}$. Clearly, if after q steps $q_j \geq \phi_j$ for $2 \leq j \leq j^*$, then:

$$Pr(Z_q \geq x) \leq Pr(Z_q^* \geq x).$$

for any x. Therefore, it suffices to find such j^* and ϕ_j such that after δn steps Z_q^* is at least $\epsilon\delta n$, this will be achieved by finding a Z_q^* that has a positive expected increase.

Because $Q > 0$, we have that:

$$\begin{aligned} Q &= \sum_j (j - \frac{3}{2})\frac{jd_j}{\sum_j jd_j} = \sum_j (j - \frac{3}{2})q_j \\ &= \sum_j (j - \frac{1}{2} - 1)q_j = \sum_j (j-1)q_j \; - \; \sum_j \frac{1}{2}q_j \\ &= \sum_{j \geq 2} (j-1)q_j - \frac{1}{2} > 0 \end{aligned}$$

Because the asymptotic degree sequence is well behaved, see definition (17). We can find a j^* such that:

$$\sum_{j \geq 2}^{j^*} (j-1)q_j > \frac{1}{2} + \epsilon'.$$

for some $\epsilon' > 0$. Therefore we can also find a sequence ϕ_j of joint probability values such that:

- $\phi_j < q_j, \quad \text{for} \quad 2 \leq j \leq j^*$.
- $\phi_1 = \phi_1 + \cdots + \phi_{j^*}$.
- $\sum_{j \geq 2} (j-1)\phi_j = \frac{1}{2} + \frac{\epsilon'}{2}$.

This gives that:

$$\sum_{j\geq 2}(j-\frac{3}{2})\phi_j = \frac{\epsilon'}{2} > 0.$$

We construct such a joint probability sequence as follows: Given that $q_j = \frac{jp_j}{K}$, for $j \geq 2$, choose any $\Delta_j > 0$ such that:

$$\frac{(jp_j - \Delta_j)}{K} \leq \phi_j < q_j$$

Taking

$$\Delta = \min_j\{\Delta_1, \Delta_2 \dots \Delta_{j^*}, \frac{1}{2}, \frac{K}{2}\},$$

then after exposing up to Δn vertices, the probability of choosing a copy of a vertex of degree $2 \leq j \leq j^*$ is at least

$$\frac{jp_j n - \Delta n}{Kn} \leq \phi_j < q_j.$$

Therefore for $0 \leq q \leq \Delta n$:

$$Pr(Z_q \geq x) \geq Pr(Z_q^* \geq x).$$

Let us now consider the variable Z_q^* with the following properties:

- $Z_0^* = 0$.
- $Z_{q+1}^* = Z_q^* + (j - \frac{3}{2})$, with probability ϕ_j for $2 \leq j \leq k^*$.

This variables has expected increase $\frac{\epsilon'}{2}$ at every step q. Therefore, after δn steps, for $delta < Delta$, its expected value is $\frac{\epsilon'}{2}\delta n$. By Chernoff's inequality, see [8], we get that:

$$Pr(Z_{\delta n}^* \leq \frac{1}{2}E(Z_{\delta n}^*)) \leq \exp\left(-\frac{E(Z_{\delta n}^*)}{4}\right)$$
$$Pr(Z_{\delta n}^* > \frac{\epsilon'}{4}\delta n) \geq 1 - \exp\left(-\frac{\epsilon'\delta n}{8}\right)$$

Therefore, if we take $\epsilon = \frac{\epsilon'}{4}$ we get that with high probability $Z_{\delta n} > \epsilon\delta n$.

□

Having shown that Z_q is very large for q large enough. We will now show that X_r also becomes very large at some point before Δn vertices have been exposed. But to do that, we need to show first that Z_q does not get too large.

Lemma 4. *If $Q > 0$ and Q finite, then there exists $\delta' \leq \frac{1}{2}$ such that for all $0 < \delta \leq \delta'$, there a.s. exists $0 < \eta < 1$ such that $Z_{\delta n} \leq \eta n$ where $\eta \leq \frac{K}{4}$. The probability of the converse is at most $(z_2)^n$ for some $0 < z_2 < 1$.*

Proof. The initial expected increase in Z_q is given by Q. We will show that even after δn steps this expected increase is not that large. We will the use an upper bound on the expected increase to bound $E(Z_q)$ and then Chernoff's inequality to bound Z_q itself.

The initial expected increase in Z_q is given by Q:

$$Q = \sum_j (j - \frac{3}{2})q_j$$
$$= \sum_j (j - \frac{3}{2})j\frac{p_j}{K} = \frac{D}{K}.$$

In the worst case, the first δn exposed vertices are all of degree $j = 1$. After δn steps the probability of choosing a copy of a vertex of degree $j \geq 2$ is:

$$q_j = j\frac{p_j}{K-\delta} \leq j\frac{p_j}{K/2} \leq 2q_j.$$

This implies that for all $q \leq \delta n$ the expected increase in Z_q is at most $2Q$. Therefore:

$$E(Z_q) \leq 2Q\delta n.$$

If we take $\delta \leq \min\{\frac{1}{2}, \frac{K}{16Q}\} = \delta'$, then by Chernoff's inequality:

$$Z_{\delta n} \leq \frac{K}{4}$$

with a probability exponential in n.

□

Remark 11. *Note that lemma (4) has no equivalent in the Molloy and Reed paper. This is put here to correct a mistake that occurred when bounding the Z_q from above in the equivalent of the proof of lemma 5 in the Molloy and Reed paper.*

Lemma 5. *If $Q > 0$, then there exists $0 < \delta'' < min(\delta', \Delta)$ for δ' as defined in lemma (4), such that for any $0 < \delta \leq \delta''$, there a.s. exist an R, $0 < R < R_{\delta n}$ such that $X_R > \gamma n$ where $\gamma = \frac{\epsilon\delta}{4}$. The probability of the converse is $(z_2)^n$ for some $0 < z_2 < 1$.*

Proof. We will bound X_r using relation (4.1.3):

$$Z_q = W_{R_q} = X_{R_q} + \frac{3}{2}Y_{R_q}$$
$$X_{R_q} = Z_q - \frac{3}{2}Y_{R_q}$$

We also have that because $X_r \geq 0$:

$$Z_q \geq \frac{3}{2}Y_{R_q}$$

Therefore if we want to bound X_{R_q}, we will have to bound the number of back edges formed Y_{R_q} from above. We will do this by counting the number of back edges formed before $X_r > \gamma n$, or $R_{\delta n}$ pairs have been formed.

At any step r, $1 \leq r < R_{\delta n}$, the probability that we form a back edge is the probability of choosing an open vertex in step 2 of algorithm (2), which is at most $\frac{X_r}{Kn-\frac{3}{2}r}$. Let us now bound R_q the number of steps required to expose q vertices:

$$R_q = q + Y_{R_q} - 1 \leq q + \frac{2}{3}Z_q - 1$$

Using the result of lemma (4), we have that $Z_q \leq \frac{K}{4}$. This gives:

$$R_q \leq q + \frac{K}{6}n - 1 \leq \delta n + \frac{K}{6}n$$

for n large enough. The probability p of forming a back-edge when $X_r \leq \gamma n$ is at most:

$$p = \frac{X_r}{Kn - \frac{3}{2}\frac{K}{6}n - \frac{3}{2}\delta n} \leq \frac{\epsilon\delta/16}{\frac{K}{2} - \frac{3}{2}\delta}$$

Consequently, the number of back edges formed has an expected value of at most:

$$E(Y_{R_q}) \leq pR_{\delta n} \leq \frac{\epsilon\delta/16}{\frac{K}{2} - \frac{3}{2}\delta}(\delta + \frac{K}{6})n$$

We would like this expected value to be less than $\frac{\epsilon\delta}{8}n$ i.e. :

$$\begin{aligned} E(Y_{R_q}) &\leq \frac{\epsilon\delta/16}{\frac{K}{2} - \frac{3}{2}\delta}(\delta + \frac{K}{6})n < \frac{\epsilon\delta}{8}n \\ \frac{\delta + \frac{K}{6}}{\frac{K}{2} - \frac{3}{2}\delta} &< 2 \\ \delta + \frac{K}{6} &< K - 3\delta \\ 4\delta &< \frac{5K}{6} \end{aligned}$$

If we take $\delta < \frac{K}{6} = \delta''$, we get by Chernoff's inequality:

$$Pr(Y_{R_q} > \frac{\epsilon\delta}{4}n) \leq (z_3)^n$$

for some $0 < z_3 < 1$.

Therefore, if for all $1 \leq r < R_{\delta n}$, $X_r \leq \frac{\epsilon\delta}{4}$, we get that with high probability:

$$Y_{R_q} \leq 2\frac{\epsilon\delta}{8}n.$$

Using inequality (A.4.1) we obtain

$$\begin{aligned} X_{R_{\delta n}} &\geq Z_{\delta n} - \frac{3}{2}Y_{R_{\delta n}} \\ &\geq \epsilon\delta - \frac{3\epsilon\delta}{8} = \frac{5\epsilon\delta}{8} > \frac{\epsilon\delta}{4}. \end{aligned}$$

□

Having shown X_r grows very large after exposing a number δn of vertices, we can show that a large number of those X_r vertices will paired within themselves to form configuration cycles, and a large number will be paired with unexposed vertices before X_r goes to zero thereby guaranteeing that after step I_δ we will expose enough new vertices to form a giant component.

Lemma 6. *If $Q > 0$, there exists constants c_1, c_2 such that the component being exposed at step $R \leq R_{\delta'' n}$, with δ'' as defined in lemma (5), has at least $c_1 n$ vertices and $c_2 n$ cycles with high probability. The probability of the converse is $(z_4)^n$ for some $0 < z_4 < 1$.*

Proof. We have shown that there exists a step $R \leq \delta'' n$ such that $X_R > \gamma n$, $0 < \gamma < 1$. We will show that with high probability $c_1 n$ of the X_R open copies will be matched with unexposed vertices and that $c_2 n$ will be matched with other open copies.

We construct a set B containing all open copies. This set has size at least γn. Also, at step $R_{\delta'' n}$, we have only exposed $\delta'' n$ vertices. From lemma (3), δ'' is at most $\frac{1}{2}$ this implies there is at least $\frac{n}{2}$ remaining vertices. We create a set A containing one copy of each of these vertices.

After R steps there at most $(Kn - \frac{3}{2}R)$ copies left to be matched. We will show that $c_1 n$ copies will be matched with members of A and $c_2 n$ copies will be matched with members of B. Our configuration building algorithm triples up, these $(Kn - \frac{3}{2}R)$ open copies of vertices

uniformly. It essentially creates $(Kn - \frac{3}{2}R)/3$ triples, with every triple created with an equal probability.

In general, given any two sets A and B which are subsets of a set C, the probability that we create a triple containing a copy from A and a copy from B from a set C is

$$\frac{|A||B||C|}{\binom{|C|}{3}}$$

where $|A|$ denotes the size of set A, and the expected number of these is:

$$\frac{|A|\ |B|\ |C|}{\binom{|C|}{3}}\ \frac{|C|}{3} \leq \frac{|A|\ |B|}{|C|}.$$

Therefore, the expected number of triples containing one copy from A and one copy from B in our configuration is:

$$\geq \frac{n/2\ \gamma n}{Kn - \frac{3}{2}R} \geq \frac{n/2\gamma n}{Kn - \delta'' n} \geq \frac{n/2\gamma/n}{Kn} \geq 2c_1 n + o(n).$$

for some constant $c_1 > 0$. The expected number of triples containing two copies of B is:

$$\geq \frac{\gamma n\ \gamma n}{Kn - \frac{3}{2}R} \geq \frac{\gamma n\ \gamma n}{Kn} = 2c_2 n + o(n).$$

for some constant $c_2 > 0$. Finally, using Chernoff's inequality we get that the number of such pairs is less than half their expected values with a probability $(z_4)^n$ for some $0 < z_4 < 1$. So, with high probability we have at least $c_1 n$ vertices with at least $c_2 n$ back-edges in the component being exposed at step R .

□

Lemma 7. *Given a configuration F as described in the first case of theorem (2), then who F has at most one component with more than $T\log(n)$ vertices for an appropriate choice of the constant T.*

Proof. We have already shown that F has at least one component of size $c_1 n$ for some constant $c_1 > 0$. We have also shown that there exist an R, $R \leq R_{c_1 n}$ such that $X_R > \gamma n$ where $\gamma = \min(\frac{\epsilon c_1}{4}, \delta'')$.

We will look at pairs of vertices (u, v) and show that the probability that u and v belong to different components of size at least $c_1 n$ and $T\log(n)$ respectively is very small. We call these components C_1 and C_2 respectively. We will show that this happens with probability $o(n^{-2})$ and therefore the expected number of such pairs is zero.

We suppose that such a pair exists and we start algorithm (2) with any copy of vertex u. If after $R \leq c_1 n$ steps of algorithm (2), we are no longer exposing C_1 then u does not lie on a component of size at least $c_1 n$, and if we have exposed a copy of v then u and v are in the same component. So we will assume neither event happens.

We modify our exploration algorithm slightly, by stopping the exploration of C_1 after R steps, and starting to explore v's component. This is legitimate because u and v are in different components and process this still produces a random configuration.

We will show that with high probability one of the vertices of C_2 will be matched with one of the X_R open copies created by the exploration of C_1. Since $X_r > \gamma n$, and the number of available copies to be matched with at any point during the exploration of C_2 is at most Kn, we get that the probability of choosing one the X_r open copies during the exploration of C_2 is at least(γ/K).

Because C_2 has at least $T\log(n)$ vertices. The probability of matching a vertex of C_2 with one of the X_R open copies from the open exploration of C_1 is at most

$$\left(1-\frac{\gamma}{K}\right)^{T\log(n)} = \left(e^{-c}\right)^{T\log(n)}.$$

for some constant c. Taking $T > 2c$ give:

$$\left(1-\frac{\gamma}{K}\right)^{T\log(n)} = o(n^{-2}).$$

Therefore the expected number of pairs (u, v) that lie on components of size at $c_1 n$ and $T\log(n)$ respectively is $o(1)$, so with high probability none exist.

□

Lemma 8. *Given a configuration F as described in the conditions of the first case of theorem (2), then whp F no components of size at most $T\log(n)$ with more than one cycle.*

Proof. We have shown that a configuration F satisfying the conditions of the first case of theorem (2) has exactly one component of size at least $c_1 n$ for some $0 < c_1 < 1$, and that all other components have size at most $T\log(n)$.

Suppose there exists one such component with at least two cycles. Let v be a vertex in such a component. We start algorithm (2) at vertex v. We will show that the probability of having two back edges is $o(n^{-1})$ and therefore no such vertices are expected to exist.

Because the size of the component of v is at most $T\log(n)$, each vertex in it has degree at most $T\log(n)$ as well. We therefore have that $X_r \leq T^2\log(n)^2$ at any step r, because the maximum number of copies of vertices consumed in the exposure of this component is the maximum number of edges, which is at most the number of vertices times the maximum degree. For the same reason we have that the component is entirely exposed in at most $T^2\log(n)^2$ steps.

The probability that a back edge is formed at any step r is at most:

$$\frac{X_r}{Kn-\frac{3}{2}r} \leq \frac{T^2\log(n)^2}{Kn-\frac{3}{2}r} \leq \frac{T^2\log(n)^2}{Kn-\frac{3}{2}T^2\log(n)^2} = o(n^{-1/4}).$$

The probability of forming at least 2 back edges is at most:

$$\binom{T\log(n)}{2}(n^{-1/4})^2 = o(n^{-1}).$$

Therefore, the expected number of vertices in components of size at most $T\log(n)$ and with more than one cycle is $o(1)$. Therefore with high probability none exist.

□

In conclusion, we have shown that for random graph that consists solely of triangles undertakes a similar qualitative behaviour as the classical random graph with a fixed edge degree sequence. We have shown that there is a criterion for the formation of the giant component namely $\sum_j jd_j(j-\frac{3}{2}) > 0$. When this sum is less than zero, the graph consists only of small components and each component has at most one cycle. Above zero we have a giant component that contains a fixed fraction of vertices of the graph and has a large number of cycles. Moreover, this component is unique and all other components are small with size at most $O(\log(n))$ and have at most one configuration cycle. This means that all small components whether before or after the threshold have a tree like structure if one ignores the small cycles formed explicitly through the triangles distribution.

Generalisation

We deduce from this result, that if a random graph that contains only triangles in its degree sequence behaves as described below and above a certain threshold, and that a random graph with a fixed single edge degree sequence behaves similarly but according to a different threshold, then a random graph that contains a mixture of edges and triangles in its degree sequence such as the one introduced by Newman [29] also behaves in the same way but for a different threshold. More specifically, we expect that below a certain point depending on the degree sequence all components are small and tree like. Above this point, we have a giant component with many cycles and all remaining components are small and tree like as well. The question that remains to be answered is how to find and prove what this threshold is.

We hope that the next section will give us a hint of what this criterion is for the case of a graph with mixed single edges and triangles degree sequence. As mentioned previously, the generating function formalism introduced by Newman [26] relies on the locally tree like structure of these random graphs to make computing certain key properties very easy. In his new model with mixed single edges and triangles he uses the same method and by doing so he is assuming the same tree like structure in this type of random graphs, assuming that we ignore the cycles formed explicitly by the triangle degree sequence. The proof presented in this chapter justifies these assumptions.

Before we move on to discussing how Newman uses the probability generating formalism to compute key properties of random graphs with fixed single edge and triangle degree sequences, we would like to discuss the possibilities of generalising the above result to structures more complex than triangles. It should be fairly intuitive to see that by simply modifying our configuration construction algorithm (2) we can adapt the Molloy and Reed proof to cliques of size 4 and beyond. Step 2 of the new algorithm would look something like

1. Repeat the following until there are no open copies left.
    - Select an open copy x_1 from S uniformly at random. Then, select another copy x_2 from S (open or non open) uniformly at random. If x_2 is not an open copy, expose its corresponding vertex by opening all its remaining copies. Remove x_2 and $(\mathbf{1/k})$th of x_1 from S.
    - Select a copy x_3 from S uniformly at random. Remove x_3 and an other $(\mathbf{1/k})$th of x_1 from S. If x_3 is a copy of an unexposed vertex, open all its remaining copies.
    -
    -
    -
    - Select a copy x_k from S uniformly at random and open its corresponding vertex. Remove x_k and an other $\mathbf{1/k}$th of x_1 from S and add the k-tuple $(x_1, \ldots, x_k)$ to F.

for the case with a random graph with a k-clique degree sequence. Furthermore, the degree sequence can specify any motif and not just cliques as long as this motif is connected. In fact, this type of degree sequences can be thought as a degree sequence of hyperedges where each hyperedge joins k vertices. We are then free to fill in every hyperedge with any motif we like.

A random graph with fixed hyperedge degree sequence can be constructed with the above algorithm. We conjecture that a random graph with a fixed k hyperedge degree sequence will have as a criterion for the formation of the giant component:

$$\sum_i i[(k-1)i - (k-1)]p_i > 0, \tag{4.3.1}$$

where p_i is the probability that random vertex is attached to i hyperedges. Note that the case of single edge and triangle degree sequences correspond to the cases $k = 2$ and $k = 3$ respectively.

We expect that such a graph will have only small components below this threshold and that these components are locally tree like if one ignores the cycles explicitly formed by the motifs that we use to fill these hyperedges. Above this threshold we expect to see a giant component with many configuration cycles and that all remaining components are small and tree like if again we ignore cycles formed explicitly.

The difficulty now remains in attempting to generalise such degree sequences even further by having a mixture of hyperedges of different size i.e. hyperedges that join groups of vertices of different size. It seems that we cannot easily adapt the Molloy and Reed proof of the classical configuration model to this type of degree sequences. This seems to be due to the fact we cannot estimate the rate of increase of open copies of vertices as we explore the components of the graph. However, we do expect the same type of behaviour i.e. locally tree like small components and a threshold above which a giant component forms. This must be the case since we can show that for any k sized hyperedge degree sequence, random graphs all behave in this same way. Therefore so must a random graph with a mixture of hyperedges of sizes at most k.

This is a very powerful result as it says that regardless of what the graph looks like locally the global structure of a random graph in a configuration model looks the same. However, it remains an open problem to determine the criterion for the formation of the giant component and prove it. We hope that the next section will give us some clues as what this criterion might be.

Chapter 5

Properties with generating functions

In this chapter, we give few examples how the generating function formalism can be applied to Newman's random graph model with clustering. The size of the percolating giant component and the distribution of the sizes of small components, have been demonstrated with few intermediate steps in his paper [29], we derive them here in full detail. The average distance between vertices in the graph and the percolation thresholds are not shown in his paper. We deriving them here for the first time by generalising the concepts of the generating function formalism of the classical configuration model.

We assume that we are given the number of single edges s_i and triangles t_i attached to each vertex i. We take p_{st} to be joint degree distribution of our graph i.e. the probability that a randomly chosen vertex is connected to s single edges and t triangles.

Definition 21. *We define the excess degree distribution q_{st} to be the probability that a vertex, reached by following a randomly selected edge in the graph, is attached to s single edges, not counting the one we used to reach it, and t triangles:*

$$q_{st} = p_{s+1,t}\frac{(s+1)}{\sum_{s,t} sp_{st}} \tag{5.0.1}$$

This is because such a vertex will have $s+1$ single edges out of $\sum_{s,t} sp_{st}$ in total. The excess degree distribution will be very useful in computations that we do later. Another that the probability of reaching a vertex of degree (s,t) by selecting a random edge and following it is equivalent to the probability reaching the same vertex, by first selecting a random vertex in the graph then randomly choosing one of its attached edges and following it to the other end.

Definition 22. *Similarly we define r_{st} to be the excess distribution with respect to triangles i.e. the probability of reaching a vertex of single degree s and triangle degree t by selecting a vertex in the graph then following the triangle to either of the two opposite vertices, not counting the triangle we used to reach it. This is given by:*

$$r_{st} = p_{s,t+1}\frac{(t+1)}{\sum_{s,t} sp_{st}}. \tag{5.0.2}$$

Definition 23. *We also define the corresponding probability generating functions*

$$g_p(x,y) = \sum_{s,t=0}^{\infty} p_{st} x^s y^t \tag{5.0.3}$$

$$g_q(x,y) = \sum_{s,t=0}^{\infty} q_{st} x^s y^t = \frac{1}{\sum_{s,t} s p_{st}} \sum_{s,t} s p_{s,t} x^{s-1} y^t = \frac{(g_p)_x(x,y)}{(g_p)_x(1,1)} \tag{5.0.4}$$

$$g_r(x,y) = \sum_{s,t=0}^{\infty} r_{st} x^s y^t = \frac{1}{\sum_{s,t} t p_{st}} \sum_{s,t} t p_{s,t} x^s y^{t-1} = \frac{(g_p)_y(x,y)}{(g_p)_y(1,1)} \tag{5.0.5}$$

By $(g_p)_x$ and $(g_p)_y$, we mean the derivative of g_p with respect to x and y respectively.

Clustering

Before we begin proving properties about this model. Let us first show that assuming a tree like structure that justifies using the generating function method, this model is indeed useful by having a non zero clustering coefficient in the limit of large graph size. Recall that the clustering coefficient is defined as:

$$\frac{3 \times \text{Number of triangle}}{\text{Number of connected triples}}$$

Let us assume the number of triangles formed by single edges is negligible. This is a fair assumption because the clustering coefficient of the configuration model with single edges is zero. This gives that the number of triangles in the graph is:

$$\sum_{s,t} t n p_{st} = n\langle t \rangle$$

If we assume that p_k is the probability that a random vertex has total degree k i.e. that $s+2t = k$. Then, the number of connected triples is:

$$\sum_k \binom{k}{2} n p_k = \frac{n}{2} \sum_k k(k-1) p_k = n \frac{\langle k^2 \rangle - \langle k \rangle}{2}$$

So for a non zero average triangle degree, we have a sparse graph with a non zero clustering coefficient in the limit of large graph size given by:

$$\frac{2\langle t \rangle}{\langle k^2 \rangle - \langle k \rangle}$$

where $k = s + 2t$.

5.1 The size of the giant component

We will derive an estimate of the size of the percolating giant component. The following derivation is justified by properties discussed in Molloy and Reed's paper on the size of the giant component [23].

We denote by ϕ the probability that a random edge is occupied in the bond percolation process. We define u to be mean the probability that a vertex reached by traversing a randomly selected edge is not in the percolating giant component. Equivalently let v be the mean probability that a vertex reached by traversing a randomly selected triangle is not a member of the giant component.

Let us first consider the case where we have no percolation (i.e. $\phi = 1$). If a vertex reached by following a randomly selected edge is not in the giant component, then all of the vertices to which it is connected (via single edges or triangles) are also not in the giant component. If this vertex is connected to s single edges and t triangles then this happens with probability $u^s v^{2t}$. Because u is the mean probability that this happens, we get the relation

$$u = \sum_{s,t} q_{st} u^s v^{2t} = g_q(u, v^2). \tag{5.1.1}$$

Similarly if a vertex reached by following a random triangle is not in the giant percolating component, then all of the vertices connected to it are not either. This yields

$$v = \sum_{s,t} r_{st} u^s v^{2t} = g_r(u, v^2). \tag{5.1.2}$$

Consequently the probability that a randomly chosen vertex is not in the giant component is

$$\sum_{s,t} p_{st} u^s v^{2t} = g_p(u, v^2) \tag{5.1.3}$$

This probability can be computed by first finding u and v using equations (5.1.1) and (5.1.1). If this can't be done analytically then it can be achieved numerically through iteration starting from an initial condition. Finally, the probability that a random vertex is in the giant component gives the fraction S of the vertices in it, which is:

$$S = 1 - g_p(u, v^2) \tag{5.1.4}$$

The idea here is that given that we know what the degree distribution p_{st} is, we can compute the generating functions g_p, g_q and g_r given in equations (5.0.3, 5.0.4, 5.0.5) to get S.

Suppose now that we have a percolating giant component. Let u be the probability that a vertex reached via a given single edge is not in the giant component. Then either this single edge is not occupied, which happens with probability $(1 - \phi)$, or it is occupied and all other vertices connected to it (via single edges or triangles) are not connected to the percolating giant component. Hence

$$u = (1 - \phi) + \phi \sum_{s,t} q_{st} u^s v^{2t} = (1 - \phi) + \phi g_q(u, v^2) \tag{5.1.5}$$

Similarly, if both vertices reached by following a random triangle are both not connected to the giant component, then either:

- The two edges leading to the opposite corners are not occupied, which happens with probability $(1 - \phi)^2$.
- Only one edge leading to the opposite corners is occupied and the other two edges are not occupied which happens with probability $2\phi(1 - \phi^2)$. In this case the one reachable vertex must have that all of its neighbours are not in th giant component.
- Any two edges in the triangle are occupied and the neighbours of both reachable vertices are not connected to the giant component. This happens with probability $[\phi^3 + 3\phi^2(1-\phi)]$.

This implies that

$$v^2 = (1-\phi)^2 + 2\phi(1-\phi^2)g_r(u,v^2) + [\phi^3 + 3\phi^2(1-\phi)]g_r^2(x,y). \tag{5.1.6}$$

Using equations (5.1.5,5.1.5) for u and v we get that the size of the giant component $nS = n(1-g_p(u,v^2))$.

Note that in the case where we have no percolation (i.e. $\phi = 1$) we get

$$u = g_q(u,v^2) \quad , \quad v = g_r(x,y), \quad , \quad S = 1 - g_p(u,v^2).$$

Which is consistent with the previous results in (5.1.1) and (5.1.2).

5.2 Small components and the phase transition

We will now try to approximate the mean size of small components (that is all components excluding the giant component). We saw from the previous chapter that for any degree sequence satisfying fairly weak conditions, a random graph almost surely has no short configuration cycles, a property that we called locally tree like. We will use this fact to estimate the size of small components using the probability generating function formalism.

Suppose we pick a random vertex v and explore its component. The shape of this component expands in a tree like fashion as in figure(5.1).

We define $h_q(z)$ to be the probability generating function for the number of vertices accessible from a vertex reached by traversing a random edge not in the giant component. Similarly, we define $h_r(z)$ to be the generating function for the number of vertices accessible from a vertex reached by traversing a random triangle, and finally $h_p(x)$ for the number of vertices accessible from a random vertex not in the giant component.

Here we will proceed similarly to the argument in [24] which we will adapt for this model. Firstly, we have that

$$h_q(z) = \sum_k P(k \text{ vertices are accessible from end of edge}) z^k$$

There are no cycles in the component, the vertices are independent. Therefore, by the power property of probability generating functions:

$$\begin{aligned} h_q(z) &= q_{0,0}z + q_{1,0}zh_q(z) + q_{0,1}zh_r(z)^2 + q_{1,1}zh_q(z)h_r(z) + q_{2,0}zh_q(z)^2 \\ &+ q_{0,2}zh_r(z)^2 + q_{2,1}zh_q(z)^2h_r(z) + q_{1,2}zh_q(z)h_r(z)^2 + \dots \\ &= zg_q(h_q(z), h_r(z)^2) \end{aligned} \tag{5.2.1}$$

Note that the extra z factor is because we count the vertex at the end of the edges as an accessible vertex in our exploration. Using a similar argument we get:

$$h_r(z) = zg_r(h_q(z), h_r(z)^2) \tag{5.2.2}$$

$$h_p(z) = zg_p(h_q(z), h_r(z)^2) \tag{5.2.3}$$

If we solve the above recursive expressions (5.2.1, 5.2.2, 5.2.3), we can compute the distribution function of the size of small components. From this we can extract the mean size as follows:

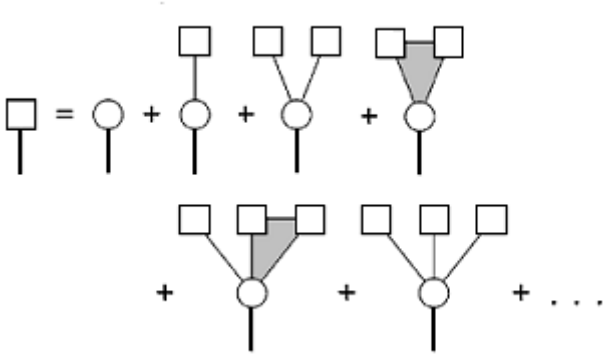

Figure 5.1: Exploiting the tree like structure to compute the size of small components using generating fucntions: The number of vertices accessible by following a randomly selected edge

$$\begin{aligned} h_p'(1) &= 1 g_p(h_q(1), h_r(1)) + (g_p)_x(h_q(1), h_r(1)) h_q'(1) \\ &\quad + (g_p)_y(h_q(1), h_r(1)) 2 h_r(1) h_r'(1) \\ &= 1 + (g_p)_x(1,1) h_q'(1) + (g_p)_y(1,1) 2 h_r'(1) \\ &= 1 + \langle s \rangle h_q'(1) + \langle t \rangle 2 h_r'(1). \end{aligned} \tag{5.2.4}$$

We can further compute $h_q'(1)$ and $h_r'(1)$:

$$h_q'(1) = 1 + \frac{(g_p)_{xx}(1,1)}{\langle s \rangle} h_q'(1) + \frac{(g_p)_{xy}(1,1)}{\langle s \rangle} 2 h_r'(1).$$
$$h_r'(1) = 1 + \frac{(g_p)_{yx}(1,1)}{\langle t \rangle} h_q'(1) + \frac{(g_p)_{yy}(1,1)}{\langle t \rangle} 2 h_r'(1).$$

where $\langle s \rangle, \langle t \rangle$ are the respective mean numbers of single edges and triangles per vertex. Note that $(g_p)_{xx} = \langle s^2 \rangle - \langle s \rangle$, $(g_p)_{yy} = \langle t^2 \rangle - \langle t \rangle$, $(g_p)_{yx} = (g_p)_{xy} = \langle st \rangle$. This gives:

$$h_q'(1)(2\langle s \rangle - \langle s^2 \rangle) = \langle s \rangle + 2\langle st \rangle h_r'(1).$$
$$h_r'(1)(3\langle t \rangle - 2\langle t^2 \rangle) = \langle t \rangle + \langle st \rangle h_q'(1).$$
$$h_q'(1)[(3\langle t \rangle - 2\langle t^2 \rangle)(2\langle s \rangle - \langle s^2 \rangle) - 2\langle st \rangle^2] = \langle s \rangle (3\langle t \rangle - 2\langle t^2 \rangle) + \langle st \rangle \langle t \rangle. \tag{5.2.5}$$
$$h_r'(1)[(3\langle t \rangle - 2\langle t^2 \rangle)(2\langle s \rangle - \langle s^2 \rangle) - 2\langle st \rangle^2] = \langle t \rangle (2\langle s \rangle - \langle s^2 \rangle) + \langle st \rangle \langle s \rangle. \tag{5.2.6}$$

Substituting (5.2.5) and (5.2.6) in equation (5.2.4) for $h_p'(1)$ we obtain an expression for the mean component size.

Most importantly, this expression has $[(3\langle t \rangle - 2\langle t^2 \rangle)(2\langle s \rangle - \langle s^2 \rangle) - 2\langle st \rangle]$ in the denominator, we conclude that the mean size of small components diverges when this equals zero, i.e. when

$$(3\langle t \rangle - 2\langle t^2 \rangle)(2\langle s \rangle - \langle s^2 \rangle) = 2\langle st \rangle^2. \tag{5.2.7}$$

The above expression forms a criterion for the position where the giant component forms. We observe that for the case of the classical configuration model (i.e. $t = 0$), the above criterion reduces to $2\langle s \rangle - \langle s^2 \rangle = 0$ which is the standard result from [22]. For the case where we have no single edges the criterion reduces to $(3\langle t \rangle - 2\langle t^2 \rangle) = 0$ which is consistent with our findings from the last chapter.

5.3 The average distance

Having shown how the tree like structure can simplify the calculation of the size of small components, we will apply the same principle again to compute the average distance between the vertices of the graph. We will achieve this by computing the distribution of the number of vertices that are a distance d away from a random vertex v.

We know that the number of vertices that are a distance 1 away (immediate neighbours) from a random vertex is generated by:

$$\begin{aligned} f(z) &= \sum_k p_k z^k = \sum_k \left(\sum_{s,t} p_{s,t} \delta_{s+2t=k} \right) z^k \\ &= \sum_{s,t} p_{s,t} z^{s+2t} = g_p(z, z^2). \end{aligned} \tag{5.3.1}$$

In order to compute the number of vertices a distance two away, we define analogously:

$$f_q(z) = g_q(z, z^2) \quad , \quad f_r(z) = g_r(z, z^2).$$

to be the generating functions for the number of neighbours of a vertex reached by traversing a random single edge or a triangle respectively. Let $f_d(z)$ be the generating function of the number of vertices that are a distance d away from a random vertex. Given that vertices are independent, we have by the power property of generating functions:

$$\begin{aligned} f_2(z) &= p_{1,0} f_q(z) + p_{0,1} f_r(z)^2 + p_{1,1} f_q(z) f_r(z)^2 + \dots \\ &= g_p(f_q(z), f_r(z)^2) = g_p(g_q(z, z^2), g_r(z, z^2)^2). \end{aligned}$$

Using a similar argument, we obtain

$$\begin{aligned} f_3(z) &= g_p(g_q(f_q(z), f_r(z)^2), g_r(f_q(z), f_r(z)^2)^2) \\ &= g_p(g_q(g_q(z, z^2), g_r(z, z^2)^2), g_r(g_q(z, z^2), g_r(z, z^2)^2)^2) \end{aligned}$$

More generally, if we define the function $F_d(x, y)$ by :

$$F_d(x, y) = \begin{cases} g_p(x, y^2) & d = 1, \\ F_{d-1}(g_q(x, y^2), g_r(x, y^2)^2) & d > 1\,. \end{cases}$$

Then taking $f_d(z) = F_d(z, z)$ we get that the generating function for the number of vertices a distance d away is:

$$F_d(z, z) = \begin{cases} f(z) & d = 1, \\ F_{d-1}(f_q(z), f_r(z)^2) & d > 1\,. \end{cases}$$

Let z_d be the mean number of vertices a distance d away from our initial vertex v. Then $z_d = f_d'(1) = F_d'(1, 1)$ and

$$F_d'(z, z) = \begin{cases} f'(z) & d = 1, \\ F_{d-1}'(f_q(z), f_r(z)^2)(f_q'(z) + 2 f_r(z) f_r'(z)) & d > 1\,. \end{cases}$$

$$z_d = \begin{cases} f'(1) & d = 1, \\ z_{d-1} \times (f_q'(1) + 2 f_r'(1)) & d > 1\,. \end{cases}$$

Hence, the average number of vertices a distance d is just a constant $a = (f_q'(1) + 2f_r'(1))$ multiple of the average of those a distance $d-1$ away. We can then write :

$$z_d = (f_q'(1) + 2f_r'(1))^{d-1} f'(1) = \left(\frac{z_2}{z_1}\right)^{d-1} z_1.$$

Now, if we approximate the average shortest path l between any two given vertices to be the distance d where we would expect to reach all the reachable vertices in the graph i.e. those in the giant component which is SN given by equation 5.1.4. In other words l is the distance where SN of the vertices are at most l away. Since $z_0 = 1$, we get the approximation

$$1 + \sum_{d=1}^{l} z_d = SN.$$

This is a geometric series, which gives

$$1 + \sum_{d=0}^{l\text{-}1} \left(\frac{z_2}{z_1}\right)^{d} z_1 = SN$$

We then get

$$z_1 \frac{(z_2/z_1)^l - 1}{(z_2/z_1) - 1} = (SN - 1) \tag{5.3.2}$$

$$l = \frac{\log\left(1 + (SN-1)(z_2 - z_1)/z_1^2\right)}{\log\left(z_2/z_1\right)} \tag{5.3.3}$$

If N is much larger than z_1 and z_2, then l behaves asymptotically like

$$l \sim \frac{\log\left(NS\right)a/z_1}{\log a}$$

Note that because the structure of the giant component is not very tree like this result is a mere rough approximation.

5.4 Percolation thresholds

We will now use the generating function formalism again to study the behaviour of the random graph under the effect of bond and site percolation. We first look at generalised site percolation, then we will look at combined uniform site and bond percolation.

In a site percolation process, we keep vertices and all edges connected to them with a certain probability, this probability could be uniform ϕ or any other function. We say that the vertex is occupied with probability ϕ.

It is often useful to have this probability as a function of the degree of the given vertex, we denote this function by ϕ_k where k is the degree of the vertex. This type of percolation is useful in simulating a targeted attack where only highly connected vertices are removed for instance, to study the effect this has on the connectivity of the network.

Suppose we have an occupation function $\phi_{s,t}$ of the degree of single vertices and triangles. We define

$$G_p(x, y) = \sum_{s,t} \phi_{s,t}\, p_{s,t}\, x^s y^t$$

to be the generating function for a vertex to have degree s, t and be occupied. We similarly define the generating functions for a vertex reached by traversing a random edge or a random triangle to have degree s, t and be occupied:

$$G_q(x, y) = \sum_{s,t} \phi_{s,t}\, q_{s,t}\, x^s y^t,$$

$$G_r(x, y) = \sum_{s,t} \phi_{s,t}\, r_{s,t}\, x^s y^t.$$

We will now look at the size of small components as we did in the previous section. We will use the approximation that these small components are tree-like and contain no cycles. As in the previous argument, suppose we traverse a random edge or triangle. Then, the generating functions for the number of accessible and occupied vertices excluding those in the giant component are:

$$H_q(z) = (1 - G_q(1, 1)) + zG_q(H_q(z), H_r(z)^2),$$
$$H_r(z) = (1 - G_r(1, 1)) + zG_r(H_q(z), H_r(z)^2).$$

respetively. Note that the first term is there because if a vertex reached is not occupied, the total number of reachable vertices is 0, and the mean probability that a random vertex is not occupied is $G_q(1, 1) = \sum_{s,t} \phi_{s,t}\, q_{s,t}$. The number of reachable vertices from a randomly selected vertex is

$$H_p(z) = (1 - G_p(1, 1)) + zG_p(H_q(z), H_r(z)^2). \tag{5.4.1}$$

Using equation (5.4.1) we can compute the expected size of small components $H'_p(1)$. The idea is that given a certain degree distribution function $p_{s,t}$ and vertex occupation function $\phi_{s,t}$, one can compute $H_q(z)$ and $H_q(z)$, if not analytically then numerically. Then we can compute H_p and the average component size $H'_p(1)$.

Finally, we can deduce the percolation threshold as the point where $H'_p(1)$ diverges, i.e the point where the expected size of components becomes infinite.

To illustrate this further, we consider the special case of uniform vertex occupation $\phi_{s,t} = \phi$. In this case, we have:

$$\begin{aligned}
G_p(x, y) = \phi g_p(x, y) \quad &, \quad H_p(z) &&= (1 - \phi) + \phi z g_p(H_q(z), H_r(z)^2) \\
G_q(x, y) = \phi g_q(x, y) \quad &, \quad H_q(z) &&= (1 - \phi) + \phi z g_q(H_q(z), H_r(z)^2) \\
G_r(x, y) = \phi g_r(x, y) \quad &, \quad H_r(z) &&= (1 - \phi) + \phi z g_r(H_q(z), H_r(z)^2)
\end{aligned}$$

Proceeding as in the calculations in section (5.2) :

$$H'_q(1)(\langle s\rangle + \phi\langle s\rangle - \phi\langle s^2\rangle) = \phi\langle s\rangle + 2\phi\langle st\rangle H'_r(1)$$
$$H'_r(1)(\langle t\rangle + 2\phi\langle t\rangle - 2\phi\langle t^2\rangle) = \phi\langle t\rangle + \phi\langle st\rangle H'_q(1)$$

Solving simultaneously we get:

$$H'_q(1) = \frac{\phi\langle s\rangle(\langle t\rangle + 2\phi\langle t\rangle - 2\phi\langle t^2\rangle) + 2\phi^2\langle st\rangle}{[(\langle s\rangle + \phi\langle s\rangle - \phi\langle s^2\rangle)(\langle t\rangle + 2\phi\langle t\rangle - 2\phi\langle t^2\rangle) - 2\phi^2\langle st\rangle^2]}$$
$$H'_r(1) = \frac{\phi\langle t\rangle(\langle s\rangle + \phi\langle s\rangle - \phi\langle s^2\rangle) + \phi^2\langle st\rangle}{(\langle s\rangle + \phi\langle s\rangle - \phi\langle s^2\rangle)(\langle t\rangle + 2\phi\langle t\rangle - 2\phi\langle t^2\rangle) - 2\phi^2\langle st\rangle^2}$$

The expected component size is

$$H_p'(1) = \phi + \phi\langle s\rangle H_q'(1) + \phi\langle t\rangle 2H_r'(1),$$

which diverges when the denominators above are zero i.e.

$$(\langle s\rangle + \phi\langle s\rangle - \phi\langle s^2\rangle)(\langle t\rangle + 2\phi\langle t\rangle - 2\phi\langle t^2\rangle) = 2\phi^2\langle st\rangle^2$$

In the case of the classical configuration model (i.e. with no triangles), the above condition gives a percolation threshold of:

$$\phi = \frac{\langle s\rangle}{\langle s^2\rangle - \langle s\rangle}.$$

which is a result previously derived by [10] using other methods.

Now consider the case of mixed uniform bond and site percolation. Denote by ϕ_s and ϕ_b the probabilities that a random vertex and edge are occupied respectively. Proceeding as before, we obtain generating functions of the degree of a vertex reached by traversing a random edge and triangle:

$$G_q(x,y) = \phi_s\phi_b g_q(x,y) \ , \quad G_r(x,y) = \phi_s(\phi_b + \phi_b^2) g_r(x,y) \ , \quad G_p(x,y) = \phi_s g_p(x,y)$$

The generating function of the number of reachable and occupied vertices in a small component:

$$\begin{aligned}
H_p(z) &= (1-\phi_s) + \phi_s z g_p(H_q(z), H_r(z)^2). \\
H_q(z) &= (1-\phi_s\phi_b) + \phi_s\phi_b z g_q(H_q(z), H_r(z)^2). \\
H_r(z)^2 &= (1-\phi_s)^2 + \phi_s^2(1-\phi_b)^2 + 2\phi_s(1-\phi_s)2(1-\phi_b)^2 \\
&\quad + [2\phi_s\phi_b(1-\phi_s\phi_b)] z g_r(H_q(z), H_r(z)^2) \\
&\quad + [\phi_s^2(\phi_b^3 + \phi^2(1-\phi_b))] z^2 g_r(H_q(z), H_r(z)^2)^2.
\end{aligned}$$

Computing the expected size of a component gives:

$$\begin{aligned}
H_p'(z) &= \phi_s + \phi_s z g_p'(H_q(z), H_r(z)^2) H_q'(z) + \phi_s z g_p'(H_q(z), H_r(z)^2) 2H_r(z) H_r'(z) \\
H_p'(1) &= \phi_s + \phi_s\langle s\rangle H_q'(1) + \phi_s\langle s\rangle 2H_r'(1) \qquad (5.4.2)
\end{aligned}$$

And the value of $H_q'(1), H_r'(1)$ are given by:

$$\begin{aligned}
H_q'(1) &= \phi_s\phi_b + \phi_s\phi_b\frac{\langle s^2\rangle - \langle s\rangle}{\langle s\rangle} H_q'(1) + \phi_s\phi_b\frac{\langle st\rangle}{\langle s\rangle} 2H_r'(1) \\
2H_r'(1) &= [2\phi_s\phi_b(1-\phi_s\phi_b)] + [2\phi_s\phi_b(1-\phi_s\phi_b)]\frac{\langle st\rangle}{\langle t\rangle} H_q'(1) \\
&\quad + [2\phi_s\phi_b(1-\phi_s\phi_b)]\frac{\langle t^2\rangle - \langle t\rangle}{\langle t\rangle} 2H_r'(1) + 2\phi_s\phi_b(3\phi_s\phi_b - 2\phi_s\phi_b^2) \\
&\quad + 2\phi_s\phi_b(3\phi_s\phi_b - 2\phi_s\phi_b^2)\frac{\langle st\rangle}{\langle t\rangle} H_q'(1) \\
&\quad + 2\phi_s\phi_b(3\phi_s\phi_b - 2\phi_s\phi_b^2)\frac{\langle t^2\rangle - \langle t\rangle}{\langle t\rangle} 2H_r'(1) \\
H_r'(1) &= [\phi_s\phi_b(1 + 2\phi_s\phi_b - 2\phi_s\phi_b^2)] \\
&\quad + [\phi_s\phi_b(1 + 2\phi_s\phi_b - 2\phi_s\phi_b^2)]\frac{\langle st\rangle}{\langle t\rangle} H_q'(1) \\
&\quad + [\phi_s\phi_b(1 + 2\phi_s\phi_b - 2\phi_s\phi_b^2)]\frac{\langle t^2\rangle - \langle t\rangle}{\langle t\rangle} 2H_r'(1)
\end{aligned}$$

Solving simultaneously we get:

$$
\begin{aligned}
&H_q'(1)(\langle s\rangle + \phi_s\phi_b(\langle s\rangle - \langle s^2\rangle)) \\
&\qquad = \langle s\rangle\phi_s\phi_b + \phi_s\phi_b\langle st\rangle 2H_r'(1) \\
&H_r'(1)(\langle t\rangle + 2[\phi_s\phi_b(1+2\phi_s\phi_b - 2\phi_s\phi_b^2)]\langle t\rangle - \langle t^2\rangle) \\
&\qquad = [\phi_s\phi_b(1+2\phi_s\phi_b - 2\phi_s\phi_b^2)]\langle t\rangle + [\phi_s\phi_b(1+2\phi_s\phi_b - 2\phi_s\phi_b^2)]\langle st\rangle H_q'(1) \\
&H_q'(1)[(\langle s\rangle + \phi_s\phi_b(\langle s\rangle - \langle s^2\rangle))(\langle t\rangle + 2[\phi_s\phi_b(1+2\phi_s\phi_b - 2\phi_s\phi_b^2)]\langle t\rangle - \langle t^2\rangle) \\
&\qquad - 2\phi_s\phi_b[\phi_s\phi_b(1+2\phi_s\phi_b - 2\phi_s\phi_b^2)]\langle st\rangle^2] \\
&\qquad = (\langle t\rangle + 2[\phi_s\phi_b(1+2\phi_s\phi_b - 2\phi_s\phi_b^2)]\langle t\rangle - \langle t^2\rangle)\langle s\rangle\phi_s\phi_b \\
&\qquad + \phi_s\phi_b\langle st\rangle 2[\phi_s\phi_b(1+2\phi_s\phi_b - 2\phi_s\phi_b^2)]\langle t\rangle \\
&H_r'(1)[(\langle s\rangle + \phi_s\phi_b(\langle s\rangle - \langle s^2\rangle))(\langle t\rangle + 2[\phi_s\phi_b(1+2\phi_s\phi_b - 2\phi_s\phi_b^2)]\langle t\rangle - \langle t^2\rangle) \\
&\qquad - 2\phi_s\phi_b[\phi_s\phi_b(1+2\phi_s\phi_b - 2\phi_s\phi_b^2)]\langle st\rangle^2] \\
&\qquad = (\langle s\rangle + \phi_s\phi_b(\langle s\rangle - \langle s^2\rangle))[\phi_s\phi_b(1+2\phi_s\phi_b - 2\phi_s\phi_b^2)]\langle t\rangle + \phi_s\phi_b\langle st\rangle\langle s\rangle\phi_s\phi_b
\end{aligned}
$$

Substituting $H_q'(1)$ and $H_r'(1)$ into $H_p'(1)$ in equation (5.4.2), we obtain an expression that diverges as before when the denominator is zero i.e.,

$$
\begin{aligned}
&(\langle s\rangle + \phi_s\phi_b(\langle s\rangle - \langle s^2\rangle))(\langle t\rangle + 2[\phi_s\phi_b(1+2\phi_s\phi_b - 2\phi_s\phi_b^2)]\langle t\rangle - \langle t^2\rangle) \\
&\qquad = 2\phi_s\phi_b[\phi_s\phi_b(1+2\phi_s\phi_b - 2\phi_s\phi_b^2)]\langle st\rangle^2,
\end{aligned}
$$

Note that for the case where we have no bond and or site percolation (i.e. $\phi_s = 1$ and or $\phi_b = 1$) the expression reduces the previous result in (5.2.7).

For example, the case where we have no site percolation (i.e. $\phi_s = 1$) we get a criterion of divergence of

$$
\begin{aligned}
&(\langle s\rangle + \phi_b\langle s\rangle - \phi_b\langle s^2\rangle)(\langle t\rangle + (2\phi_b + 4\phi_b^2 - 4\phi_b^3)\langle t\rangle - (2\phi_b + 4\phi_b^2 - 4\phi_b^3)\langle t^2\rangle) \\
&\qquad = (2\phi_b + 4\phi_b^2 - 4\phi_b^3)\langle st\rangle^2.
\end{aligned}
$$

In the case of the classical configuration model with no triangles, this reduces to

$$
\langle s\rangle + \phi_b\langle s\rangle - \phi_b\langle s^2\rangle = 0
$$

Solving this gives a bond percolation threshold of $\phi_b = \frac{\langle s\rangle}{\langle s^2\rangle - \langle s\rangle}$.

Chapter 6

Criticism and future models

In this chapter we will discuss the limitation of Newman's model in terms of applications, and how we can improve on them by generalising it to models of several degree distributions and higher order motifs.

6.1 Limited clustering

We have shown previously that for a sparse degree sequence, a Newman random graph has a non zero clustering coefficient in the limit of large graph size.

To maximise the clustering in Newman's model, we would like our vertices to be connected to triangles only. In this case, the degree of a vertex is two times the number of triangles i.e $\langle k \rangle = 2\langle t \rangle$. The total clustering coefficient is by given by the previously derived formula:

$$\frac{2\langle (k/2) \rangle}{\langle k^2 \rangle - \langle k \rangle} = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}$$

.

So when $\langle k^2 \rangle$ is very close to $\langle k \rangle$, we have a high clustering coefficient but how does this restrict the degree of vertices ?

Let us now look more closely at the local clustering coefficient defined previously:

$$C_i = \frac{3 \times \text{number of triangles connected to vertex i}}{\text{number of connected triples around vertex i}}$$

We will consider C_k, the local clustering coefficient as a function of a vertex of degree k. In this case a vertex of degree k is connected to $k/2$ triangles, this gives a local clustering coefficient as a function of degree of:

$$C_k = \frac{k/2}{\binom{k}{2}} = \frac{1}{k-1}.$$

So the maximum clustering coefficient of a vertex of degree k is $(1/k - 1)$.

6.2 Gleeson's model

In order to provide some contrast to Newman's model, we briefly introduce another model also very recently published that has non zero clustering in the limit of large graph size and that is tractable. Gleeon's model [14], is very similar to Newman's model in that it is a generalisation of the classical configuration by specifying a joint degree sequence for single edges and other higher order motifs.

In this model, vertices are connected to singles edges and/or to cliques of variables size c. However, every vertex can only be part of one clique. We are given a joint degree distribution $\gamma_{k,c}$ that specifies the probability that a randomly chosen vertex has degree k and is part of a clique of c vertices. This illustrated by figure (6.1).

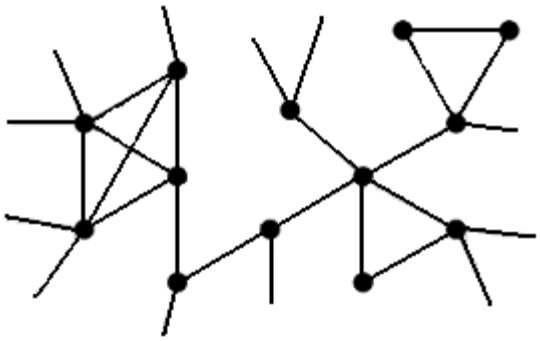

Figure 6.1: Single edges and cliques, Gleeson's Model.

If a node is is not part of any clique then it is said to be a member of a 1-clique. By having cliques as large as we want we can tune the clustering coefficient. One way to visualise this random graph is that cliques are super-nodes connected to each other via singles edges, see [14] and figure (6.1). Viewed this way, the graph of super-nodes is simply a graph constructed using the classical configuration model. It therefore has the same tree like structure and zero clustering.

Let us try to compute the global clustering coefficient C. Given that the graph of super-nodes is tree like, there are no triangles except inside the cliques. The clustering coefficient is therefore:

$$\frac{3n\sum_{ck}\gamma_{ck}\binom{c}{3}}{n\sum_{kc}\gamma_{kc}\binom{k}{2}} = \frac{\langle c^3\rangle - 3\langle c^2\rangle + 2\langle c\rangle}{\langle k^2\rangle - \langle k\rangle}$$

Suppose that a vertex i has degree k and is connected to a c clique. The local clustering coefficient therefore is:

$$C_k = \frac{\binom{c-1}{2}}{\binom{k}{2}}$$

We can see that contrarily to Newman's model, the clustering coefficient is not as restricted by the degree k. We can in fact tune it to be high as 1 up to any degree k' by taking $c = k' + 1$, as a vertex with degree k can take part in a cliques of size up to $k + 1$.

6.2.1 Key properties

In his paper [14], Gleeson prefers a cascade method to prove the size of a percolating giant component for a random graph with degree sequence γ_{kc}. We have attempted to apply the generating function method to calculate certain properties like the size of small components. However this does not seem possible because the lack of independence between vertices of the same cliques.

Indeed, one can easily compute the excess distribution of single edges as is done in Newman's model, but one cannot compute an excess distribution for cliques because if we select a random vertex x and traverse the clique to reach another vertex y, then y must have the same c clique size as x.

One may also be tempted to reduce this model to the classical configuration model by looking at it as a graph of super-nodes. To do this we will need to compute the degree distribution of the graph of the super-nodes from the distribution γ_{kc} but it is impossible to compute such a fixed degree distribution because the degree distribution of the super nodes will vary from one configuration to another. All configurations are equally likely. We could however compute an expected degree distribution for the super-nodes in certain cases. The easiest of such cases is when the number of single edges and clique sizes are independent.

6.3 Generalisations and open problems

We will now analyse the pros and cons of these two models and some generalisations that would improve these. Both of the models that we have seen so far are generalisations of the classical configuration model. In other words, the random graph with a fixed degree sequence are a special case of both models. They also both have provable non zero clustering coefficients in the limit of large graph size for sparse graphs.

Newman's model provides a intuitive way to create a non zero clustering coefficient by placing triangles directly on vertices. It also offers flexibility by specifying exactly the number of singles edges and triangles each vertex has. The main drawback is that for a given degree k the local clustering coefficient is at most $1/(k-1)$.

Gleeson's model improves on the clustering limitation, by allowing cliques of variable size and not just triangles. It also improves on the local clustering by allowing any value of the local clustering coefficient regardless of the degree. However it is not very flexible in that it allows only one clique per vertex. It is however flexible in tuning the clustering coefficient as Gleeson shows with an example in his paper [14]. One can choose an appropriate γ_{kc} distribution to obtain any distribution for the clustering coefficient C_k as a function of the degree.

6.3.1 Tractability

We have shown through chapter 4, that the Molloy and Reed proof can be adapted and generalised to any type of random graphs with a fixed degree distribution of fixed size motif. We concluded from this that for any random graph with fixed distribution of these motifs will behave qualitatively like the classical configuration model. It will have a threshold for appearance of a unique multi-cyclic giant component, otherwise it has only tree like small components. Furthermore, we argued that the same qualitative behaviour applies to any fixed distribution of mixed motifs like Newman's mixed single edges and triangles model.

This result implies that both a Newman or a Gleeson random graph has a locally tree like structure in its small components and has a threshold for the formation of the giant component. This property can be exploited by using methods like probability generating function in the case of Newman's model and tree cascades in the case of Gleeson to compute certain key properties of these graphs.

We comment however, that Newman's model seems easier to work with and that the generating functions methods seems more powerful in deriving many key properties for any degree distribution. In our attempts to apply this method to Gleeson's model, we believe that this is not possible because the size of cliques is variable and vertices within a clique are not independent which prohibits the formulation of an excess degree distribution for cliques as Newman does for his new model and the classical configuration model.

6.3.2 Applications

In terms of applications, Newman's model is very good as it provides a natural and intuitive way to implement clustering. The degree distribution of triangles can easily be measured in networks

taken from the real world and then implemented into the model. Placing triangles directly onto vertices is also an intuitively justifiable way to incorporate clustering into a network under the triadic closure definition of the clustering coefficient.

On the other hand, the maximum local clustering coefficient is bounded above by $1/(k-1)$. This is very low and unlike real world networks who tend to have much higher clustering. Social networks for examples tend to have high degrees, and clustering coefficients of the order of tens of percents simultaneously, see [25]. This seriously limits the prospects of applications of this model.

Gleeson's model has a tunable clustering coefficient, which as demonstrated by Gleeson in his paper [14] can be easily made to match any distribution of the clustering coefficient C_k and therefore model many of real life networks. However, it is quite artificial in its restriction to have only one clique per vertex. It is very hard to justify for instance why an individual in a social network is a member of no more than one community.

6.4 Generalisations

Given the success of the two models presented in this dissertation in solving a long standing problem in the study of network by creating simple, flexible and tractable models of random graphs with provable non zero clustering in the limit of large graphs size, it is very exciting to ask the question how far can we generalise these two models to make them more flexible and powerful whilst still tractable.

6.4.1 Newman's model

The main con of this model is that it has limited local clustering for a given degree which is much lower than what we find in real world networks. A simple way to improve on the local clustering coefficient is to have higher order cliques in our degree distribution. Suppose we generalise this model by having a joint degree sequence $p_{c,s}$ for the number of single edges s and cliques of size c attached to each vertex. Then, for a vertex of degree k the local clustering coefficient is:

$$C_k = \frac{k/(c-1)\binom{c-1}{2}}{\binom{k}{2}} = \frac{c-2}{k-1}$$

So by choosing a value of c large enough we can have a large clustering coefficient up to a certain degree k. However, having a distribution of cliques of size c is very artificial and cannot be justified in terms of applications.

It would be natural to specify a distribution of different sized cliques. For example, we could specify a distribution p_{stuv} which represents the probability that a random vertex is connected to s single edges, t triangles, u 4 cliques and v 5 cliques. This type of distribution can be easily measured from a real world network and is intuitively justifiable in a real life network.

Of course, we do not have to limit ourselves to cliques but we can have any connected motifs. Ultimately we can have a degree distribution of the form $p_{k_1 \dots k_m}$ specifying the number k_i of each connected motif i. This type of model will still be tractable in the usual way using generating functions at the expense of heavier calculations: We would have m different excess degree distributions and m generating functions. The maximum clustering would depend on the motif with largest number of triangles.

6.4.2 Gleeson's model

The most obvious way to generalise this model would be to allow vertices to take part in multiple cliques. From the perspective of applications it is natural that an entity belongs to a variable number of communities. It is also very natural to assume that this variable number can be

bounded by a certain value m. We would obtain a distribution of the form $\gamma_{kc_1...c_m}$, which we could define in a way such that if a node is a member of less than m cliques the corresponding c_i values will have values 1. Furthermore, we do not have to be restricted to cliques, but we can use any connected motifs that grow in a specified fashion. For example c_i can specify a cycle of size c_i and c_j can specify a binary tree of a certain size etc.

Again this generalised model will still be solvable using the same tree cascade method used by Gleeson [14] at the expense of more calculations.

6.4.3 Molloy and Reed proofs

As stated previously, we have omitted an adapted proof for the Newman model, which was aimed to show that this model has the same qualitative behaviour as the classical configuration model. Instead we presented an adapted version of the special case where we have only triangles in our degree distribution. We argued that consequently this implied that the mixed single edges and triangles model behaved in the same away because it was kind of sandwished between the classical configuration model and the triangle configuration model. Because of this, one might argue that the first proof is not required. However, these proofs not only provide us with qualitative descriptions of the behaviour of the graph but they also specify criteria for the point where the giant components forms. We also think that if one can succeed in proving this result for Newman's model, we can very easily adapt it to a more general mix of different motifs.

As we argued in chapter 4, we can adapt the Molloy and Reed proof to any random graph with fixed distribution for any fixed size connected motif and show a result analogeous to theorem (2). We described how we could adapt the configuration construction algorithm to construct a random graph with a fixed hyperedge distribution as long as the hyperedges include the same number of vertices. We can then fill these hyperedges with any connected motif we like. The difficulty seems to arise when we have hyperedges with different sizes (include different numbers of vertices). More specifically, we find it hard to estimate the initial rate of increase of the number of open vertices as we explore component. This rate of increase seems to define, in all the cases we could solve, the criterion of the formation of the giant component.

We saw however, using Newman's generating functions method that computing this criterion is possible. These criterions must be true since we argued that for any distribution of mixed single edges and triangles, the components of the graph are tree like and therefore the generating function results provide good approximations. All this provides good reasons to believe that adapting the Molloy and Reed proof to Newman's model of single edges and triangles is a possible target.

We believe that once we do this for the simplest case of single edges and triangles we could adapt it to the most general form of Newman's model described in this chapter, where we have random graphs of fixed mixed distribution of any finitely sized connected motifs.

Finally, if we would like to adapt the Molloy and Reed proof for the Gleeson model, things seem less obvious. There is an added difficulty here in that hyperedges have an arbitrary size making the calculation of the expected increase even less straightforward. Furthermore, the key properties calculated by Gleeson in his paper does not include the critical point where the giant component forms.

Conclusion

In this dissertation, we were motivated by the publication of two very recent and promising papers, that claimed they had solved one of the long standing problems in network theory, by presenting two tractable random graph models with a non zero clustering coefficient in the limit of large graph size. These papers were written by physicists who used heuristic methods and approximations that were justified by more rigorous work done by other people like Molloy and Reed [22] but for different random graph models. We set out to study these models and make the results of these authors more rigorous by providing proofs for the global structure and qualitative behaviour of graphs in their new models.

In the first few chapters we presented the background theory of this field. These chapters reflect the learning curve of this dissertation. Network theory is a very vast topic that is discussed in different types of literature written by people from different disciplines who use different methods, some more rigorous than others.

Chapters 4, 5 and 6 represent the more creative part of this dissertation. We gave a proof for the qualitative behaviour of a random graph with a fixed triangle sequence. We defined a criterion of the formation of the giant component. We showed that all small components were tree like and we also gave bounds on the size of these components. In the process of this, we corrected some mistakes found in the paper by Molloy and Reed [22]. We then showed how this proof can be generalised to any random graph model with a fixed hyperedges degree distribution that could then be used to form random graphs with fixed motif distributions. We Argued that this result implied that a random graph with mixed single edge and triangle distribution must have the same qualitative behaviour as it is sandwiched between the classical configuration model and the triangles model that we looked at. Therefore, the same also applied to any random graph with a mixed distribution of different size motifs.

We then claimed that we could use these results to justify the use of the generating function method for computing key properties for Newman's model as he does in his paper [29]. We derived some key results that were briefly discussed in his paper as well as some new results. We showed that the results derived using these methods were consistent with results derived by other people using more rigorous methods. Most importantly, the criterion of the formation of the giant component in a graph with only triangles was consistent with the one we show in our proof of theorem (2).

We then looked at the strengths and weaknesses of the models that we presented. We proposed many generalisations of these models and discussed their relevance to applications and how we could go about computing their properties.

Having achieved all this, we feel that there are many more problems that, given more time, we would have liked to attempt. The most important of these are:

- Prove a result equivalent to Janson's [18], on the condition required that the probability that a random configuration creates a simple graph is positive. We would like to derive and prove a similar result for the more general case of a configuration of hyperedges. This is an important intermediate result for proofs on generalised configuration models.

- Look at and Molloy and Reed's second paper [23] on the size of the giant component. Generalising their results in this paper as we did here can help justify the use the generating function method for properties of the giant component.

- Research more properties of the configuration model that were shown using more rigorous methods and check for their consistency with the predictions of the generating functions method.

- Compute more key properties of Gleeson's model and check they are consistent with previous results.

- Study the generalised models proposed in chapter 6 in more detail to see how the choice of different types of motifs and degree sequences affect their key properties as we did with the clustering coefficient.
- Prove the result that we could not show here, namely an equivalent of theorem (2) for Newman's model of joint degree distribution of single edges and triangles. We believe that this is possible if we use Newman's criterion for the formation of the giant derived in chapter 5, as a basis for future investigations. Having achieved this, we would like to carry on and do the same for the general case of models with hyperedges degree sequences where hyperedges have different size.

We believe that Newman's joint edge and triangle degree distributions and Gleeson's degree and clique participation distributions, are the first of few models that will have very powerful applications in real life networks that exhibit high clustering such as social networks. The generalisations that we discussed here are only few examples of further work that could be done in this area. Newman and Gleeson's papers have opened new doors in the research of network theory and random graphs, a subject that continues to be very dynamic and exciting.

Appendix A

A proof for the new model

The following is the omitted proof of Newman's model of single edge and triangle degree configuration. We have decided to put this proof here after discovering that the criterion of the formation of the giant component, derived using Newman's generating functions (5.2.7) is not equivalent to the result that we arrive at here (A.1.2). We did not have the time to investigate what mistakes may have been made, but we suspect that this is probably due to our approximation that on average we select a blue copy with probability s and a red copy with probability t see algorithm (3) and definitions (31).

We will start by defining a few concepts, we will then state our result in the form of a theorem, which we will split into several lemmas which we will then prove.

Definition 24. *A joint degree sequence is a sequence of pairs of non negative integers $(s,t)_1 \dots (s,t)_n$ that represents the number of single edges s and triangles t attached to a vertex i. We will usually denote this by (s_i, t_i).*

Definition 25. *We say that a joint degree sequence is feasible if the set of all possible graphs with that sequence is non-empty.*

Definition 26. *An asymptotic joint degree sequence is a sequence of pairs of integer valued function $(s,t)_1(n), (s,t)_2(n) \dots$, such that for a fixed graph size n we obtain a fixed joint degree sequence (s_i, t_i). Note that this definition is not similar to the degree sequence of the classical model in 9.*

Definition 27. *Throughout the following chapters, we will use the terms*

- *Joint degree to refer to the pair (s_i, t_i) denoting the number of single edges s_i and triangles t_i attached to a vertex i.*
- *Sum degree to denote the value $(s_i + t_i)$ for a vertex i with joint degree (s_i, t_i), we denote this value by $deg(i)$.*
- *Total degree to denote the actual number of edges attached to a vertex. This is given by $s_i + 2t_i$ for a vertex with joint degree (s_i, t_i).*

A.1 A configuration model with triangles

Suppose we are given a joint degree sequence $(s_1, t_1), (s_2, t_2), \dots$ representing the number of single edges and triangles respectively for a each vertex i in a graph of n vertices.

Using this sequence, we construct the matrix $d_{i,j}$ or $d(i,j)$, where each entry represents the number of vertices in the graph with exactly i single edges and j triangles.

Firstly, We Construct a set S of red and blue copies of vertices for our graph, by creating s *blue* copies for every vertex with s single edges and t red copies for every vertex with t triangles attached to it. In total we have $s_i + t_i$ copies for each vertex i.

Definition 28. *A random configuration is a partition of S that consists of a set of pairings of the blue copies of S and triples of the red copies of S. this partition is selected uniformly at random from the set of all possible partitions.*

Definition 29. *An configuration cycle is a cycle that is not created explicitly by a triangle in the degree sequence.*

We will construct a random graph G with the above degree sequences by constructing a *random configuration* using the following algorithm.

Algorithm 3. *We construct our configuration F by pairing the blue copies and tripling the red copies of the set S. We say that a vertex is exposed if any of its copies has been added to F, and we say that the copies of an exposed vertex that remains in S are open.*

Repeat the following until S is empty:

1. *Expose a random vertex v in G by selecting a random element or copy in S then exposing all the remaining copies of the same vertex.*
2. *Select an open copy x from S uniformly at random.*
   - *If this copy x is blue we uniformly select another blue element y from S and pair it with x, we add the pair to F and delete it from S. If the vertex corresponding to y is not exposed, we expose it and open all its remaining copies.*
   - *If this copy is a red copy, We uniformly select one more copy y choosing uniformly from L, we say we pair y with half of x, if the vertex corresponding to y is unexposed we open all its other copies, we remove y and x half of x from S.*
   - *If the copy selected is half a red copy, we pair it with a randomly selected red copy from S and expose its other copies. We remove these from S and add the triple composed of x, y and the copy paired with the other half of x to F.*

   Repeat step 2 as long as there any open copies left. otherwise go to step 1.

We can see that using this algorithm, we construct any configuration with the specified joint degree sequences uniformly at random from the set of all possibilities. The action of pairing two blue copies corresponds to connecting two vertices with a single edge. The action of tripling three red copies corresponds to connecting three vertices in a triangle. Hence, the algorithm is exposing the components of G one at a time, a component is fully exposed when there are no more open copies and a new component is started we go back to step 1. Note also that in step 1, the vertex did not have to be selected at random, it could be any vertex whose component we would like to expose.

Of course, the above algorithm essentially constructs a multi-graph, but for the case of the classical configuration model Janson [17] showed that for the classical model, given certain conditions, there is a positive probability that the graph is simple. Although we do not show an equivalent results here, in what follows we condition on the fact that graph we obtain is simple.

Definition 30. *We say that a joint degree sequence is sparse if the sum of all sum degrees of the vertices of the graph is linear in the size of the graph*

$$\sum_{i\geq 0} s_i + t_i = \sum_{i\geq 0} (i+j)d_{i,j} = Kn + o(n).$$

Definition 31. *We say that a joint degree sequence is well-behaved if it is feasible and there exist constants $p_{i,j}, s, t$ such that*

1. $$\lim_{n\to\infty} \frac{d_{i,j}(n)}{n} = p_{i,j}.$$

2. $$\lim_{n\to\infty} \frac{\sum_{i,j} i d_{i,j}(n)}{\sum_{i,j}(i+j)d_{i,j}(n)} = \lim_{n\to\infty} \frac{\sum_{i,j} i p_{i,j}}{\sum_{i,j}(i+j)p_{i,j}} = s.$$

3. $$\lim_{n\to\infty} \frac{\sum_{i,j} j d_{i,j}(n)}{\sum_{i,j}(i+j)d_{i,j}(n)} = \lim_{n\to\infty} \frac{\sum_{i,j} j p_{i,j}}{\sum_{i,j}(i+j)p_{i,j}} = t.$$

4. $$\lim_{n\to\infty} \sum_{i,j}(i+j)(i+j-2s-\frac{3}{2}t)\frac{d_{i,j}}{n} = \sum_{i,j}(i+j)(i+j-2s-\frac{3}{2}t)p_{i,j}.$$

Note that s and t represent the ratio of the blue and red copies in S respectively. We define

$$D = \sum_{i,j}(i+j)(i+j-2s-\frac{3}{2}t)p_{i,j} \tag{A.1.1}$$

$$Q = \frac{\sum_{i,j}(i+j)d_{i,j}(i+j-2s-\frac{3}{2}t)}{\sum_{i,j}(i+j)d_{i,j}} = \frac{D}{K} \tag{A.1.2}$$

Theorem 3. *Let $(s_1,t_1)(n),(s_2,t_2)(n),\dots$ be a sparse, well-behaved asymptotic joint degree sequence such that the probability that a random configuration with this sequence constructs a simple graph is positive. Let G be a graph with n vertices and the above joint degree sequence chosen uniformly at random from the set of all graphs with such sequence, then*

- *If $D > 0$ and if Q is finite and if there exists a constant $0 < \beta < 1$ such the number of vertices attached to both single edges and triangles is at least βn. Then, there exist constants $c_1, c_2, c_3 > 0$ such that G a.s. has one component with at least $c_1 n$ vertices and $c_2 n$ configuration cycles. Furthermore, G a.s. has exactly no other component with size greater than $c_3 \log(n)$ and more than one configuration cycle.*
- *If $D < 0$ and there exists an $\epsilon > 0$ and a function $w(n)$ such that if $0 \le w(n) \le n^{1/8-\epsilon}$ and the maximum sum degree of the sequence is at most $w(n)$ for all n. Then there exists a constant R such that G almost surely has no component with size greater than $Rw(n)^2 \log n$ vertices and more than one configuration cycle.*

A.2 The rate of growth

We motivate the main idea behind the proof as follows: If the initial rate of increase of open copies is positive, then we are likely to expose many vertices and form a giant component. If it is negative the number of open copies runs to zero and we expose a small component. Let us first we define few variables that will be useful later.

Definition 32. *We define the following variables :*

- *Let X_r be the number of open copies after the r^{th} pair has been formed. Note that a triple here is counted as two pairings.*
- *Let C_r be the number of components fully or partially exposed when the r^{th} pair has been exposed, again counting a triple as two pairs.*
- *We say that a back-edge has been formed when we pair an open copy of S with another open copy of S in step 2. This in fact corresponds to forming a configuration cycle.*
- *Let Y_r be the number of back-edges formed by pairing two open blue copies, when the rth pair has been exposed.*
- *Let Y'_r be the number of back edges formed by pairing an open red copy in step 2 with an other (half) open red copy.*

We will now motivate the remainder of our proof by looking at the initial rate of increase of open copies X_r.

Remark 12. *Let us compute the expected increase in number of red copies. We will do this by conditioning on the colour of the first vertex chosen.*

- If the first open copy selected in step two is a blue copy, the expected increase is

$$\sum_{i,j} \frac{d_{i,j}}{\sum_{i,j} d_{i,j}} i(i+j-2).$$

  This is because every blue copy in S is selected with a probability $(id_{i,j}/\sum_{i,j} id_{i,j})$, and by doing so we add $(i+j)$ new open copies and remove two.

- If the open copy selected is a red copy, then we expose one new vertex and the expected initial increase is

$$= \sum_{i,j} \frac{d_{i,j}}{\sum_{i,j} jd_{i,j}} j \left(i + j - \frac{3}{2} \right).$$

Recall from definition 31 that

$$s = \frac{\sum id_{i,j}}{\sum d_{i,j}(i+j)} \quad , \quad t = \frac{\sum jd_{i,j}}{\sum d_{i,j}(i+j)}.$$

Initially, these constants correspond to the probability that a randomly chosen copy in S is either a blue or red copy. Now, given that the vertex chosen in step 2 of algorithm 2 is chosen uniformly at random out of all copies in S. By conditioning on the colour of this first vertex, we obtain an expected number of open copies given by

$$
\begin{aligned}
&s\sum_{i,j}\frac{d_{i,j}}{\sum_{i,j} id_{i,j}}i(i+j-2) \;+\; t\sum_{i,j}\frac{d_{i,j}}{\sum_{i,j} jd_{i,j}}j(i+j-\frac{3}{2})\\
&=\sum_{i,j}\frac{d_{i,j}}{\sum_{i,j} d_{i,j}(i+j)}i(i+j-2) \;+\; \sum_{i,j}\frac{d_{i,j}}{\sum_{i,j} d_{i,j}(i+j)}j(i+j-\frac{3}{2})\\
&=\sum_{i,j}\frac{d_{i,j}}{\sum_{i,j} d_{i,j}(i+j)}i(i+j-2)+j(i+j-\frac{3}{2})\\
&=\sum_{i,j}\frac{d_{i,j}}{\sum_{i,j} d_{i,j}(i+j)}(i+j)(i+j)-2\sum_{i,j}\frac{d_{i,j}i}{\sum_{i,j} d_{i,j}(i+j)}\\
&\quad-\frac{3}{2}\sum_{i,j}\frac{d_{i,j}j}{\sum_{i,j} d_{i,j}(i+j)}.
\end{aligned}
$$

Using the fact that $\sum id_{i,j} = s\sum d_{i,j}(i+j)$ and $\sum jd_{i,j} = t\sum d_{i,j}(i+j)$, we get

$$\sum_{i,j}\frac{d_{i,j}(i+j)}{\sum_{i,j} d_{i,j}(i+j)}(i+j-2s-\frac{3}{2}t) \tag{A.2.1}$$

Definition 33. *We define the following useful variables :*

- *We define the variable Z_q to be the sum of $(i+j-2s-\frac{3}{2}t)$ over the first q exposed vertices.*
- *We also define the analogous variable W_r to be the sum over $(i+j-2s-\frac{3}{2}t)$ over all vertices exposed by the time rth pair has been exposed, counting a triple as two pairs.*

Remark 13. *The reason we introduce Z_q is that it has the same rate of increase as X_r but behaves much more nicely in that it only increases by $(i+j-2s-\frac{3}{2}t)$ every time a vertex with i single edges and j triangles is exposed. Hence, it is easy to put a bound on it's expected value when a fixed number of vertices have been exposed as we shall see later.*

We now relate all the variables defined previously. We define the variable R_q to be the number of pairs exposed by the time we expose the qth vertex i.e. $W_{R_q} = Z_q$.

Remark 14. *Note that X_r is (roughly) the same as W_r except when we form a back edge. In which case X_r decreases, note also that when we form our first pair or triple, W_r is already less than X_r by either $\frac{3}{2}$ or 2. Hence we obtain*

$$W_r = X_r + 2Y_r + \frac{3}{2}Y'_r - (2s+\frac{3}{2}t)C_r. \tag{A.2.2}$$

Remark 15. *We can also relate W_r to Z_q. If no back edges are formed we would have exposed $R_r = r$ vertices. Consequently we would get $W_r = Z_r$, but given that we get some back edges we get $r = R_r - Y_r - Y'_r$, so*

$$W_r = Z_{(r-Y_r-Y'_r)}. \tag{A.2.3}$$

Remark 16. *Z_r changes by at most $1-2s-\frac{3}{2}t = -(s+\frac{t}{2}) > -1$ every time a vertex is exposed. This happens when we expose a vertex with sum degree 1 (i.e $i+j=1$)*

$$\begin{aligned} Z_r &\geq Z_{(r-Y_r-Y'_{r'})} - (s+\frac{t}{2})(Y_r+Y'_r) \\ &= W_r - (s+\frac{t}{2})(Y_r+Y'_r) \\ &= X_r + (\frac{t}{2}+1)Y_r + \frac{t}{2}Y'r - (2s+\frac{3}{2}t)C_r \\ &\geq X_r - 2. \end{aligned}$$

A.3 Small components

We now show that if the conditions of the second case of theorem (3) are satisfied, the graph has no components of size larger than $\alpha = Sw(n)^2 log(n)$ vertices.

Lemma 9. *Let G be a graph that satisfies the conditions of the second case of the theorem. Let v be any vertex then the probability that v lies in a component of size $\alpha = Sw(n)^2 log(n)$ is less than n^{-2}.*

Proof. Suppose that we start our algorithm by choosing v at step 1. We have that

$$Q = \frac{D}{K} = \sum_{i,j} \frac{d_{i,j}(i+j)}{\sum_{i,j} d_{i,j}(i+j)}(i+j-2s-\frac{3}{2}t) < 0.$$

The probability that a given component has size at least α is at most the probability that $X_\alpha > 0$, which is consequently at most the probability that $Z_r > -2$ from remark (16). This is because if $X_r = 0$ then we would have exposed the whole component.

Initially the rate of increase of Z_r is

$$\sum_{i,j} \frac{d_{i,j}(i+j)}{\sum_{i,j} d_{i,j}(i+j)}(i+j-2s-\frac{3}{2}t).$$

After exposing $q \leq \alpha$ vertices, the rate of growth of Z is highest if the first q vertices that were exposed have sum degree 1(i.e. $i+j=1$). Hence the rate of increase of Z_q is at most

$$\frac{-(s+\frac{t}{2})(d_1-q) + \sum_{i+j\geq 2} d_{i,j}(i+j)(i+j-2s-\frac{3}{2}t)}{(d_1-q)+\sum_{i+j\geq 2} d_{i,j}(i+j)}. \tag{A.3.1}$$

where d_1 refers to the number of vertices connected to either one edge or one triangle ($i+j=1$), (A.3.1) is at most

$$\begin{aligned} &\leq \frac{\sum_{i,j} d_{i,j}(i+j)(i+j-2s-\frac{3}{2}t)}{\sum_{i+j\geq 2} d_{i,j}(i+j) \quad -q} + \frac{(s+\frac{t}{2})q}{\sum_{i+j\geq 2} d_{i,j}(i+j)-q} \\ &\leq \frac{\sum_{i,j} d_{i,j}(i+j)(i+j-2s-\frac{3}{2}t)}{\sum_{i+j\geq 2} d_{i,j}(i+j)} + \frac{(s+\frac{t}{2})q}{\sum_{i+j\geq 2} d_{i,j}(i+j)-q}. \end{aligned}$$

Because $q \leq \alpha = o(n)$ and $\sum_{i+j\geq 2} d_{i,j}(i+j) \sim Kn = \theta(n)$, we get

$$\leq Q + o(1) \leq \frac{3Q}{4} < 0.$$

The expected increase in Z_q is still negative, indicating that the process should die out quickly. Given that the degree of the the first chosen vertex v is at most $w(n)$, we get that after α vertices the **expected value** of Z_α is at most

$$\text{Initial value} + (\text{Rate } \times \alpha) \leq (3Q/4)\alpha + w(n).$$

Because $\alpha = Sw(n)^2 log(n)$, for n large enough it follows that

$$\frac{3Q}{4}\alpha + w(n) \leq \frac{Q}{2}\alpha.$$

We now introduce an important result known as Azuma's inequality that will help bound the probability of Z deviating too far from its mean.

Azuma's inequality

Let $X_0, \ldots, X_n$ be a martingale with $|X_i - X_{i-1}| \leq 1$, for all $0 \leq i < n$, with Let $\lambda > 0$ it follows that

$$Pr(|X_n| > \lambda\sqrt{n}) < e^{-\lambda^2/2}.$$

Azuma's inequality yields the following standard corollary.

Corollary 2. *Let $\Sigma = \Sigma_1, \ldots, \Sigma_n$ be a sequence of random events. Let $f(\Sigma) = f(\Sigma_1, \Sigma_2, \ldots, \Sigma_n)$ be a random variable defined over these events. Then if $E(f|\Sigma_1, \Sigma_2, \ldots, \Sigma_i)$ is $c-Lipshtiz$, that is if there exists constants c_i , $c = (c_1, \ldots, c_n)$ such that for all i :*

$$max|E(f(\Sigma)|\Sigma_1, \ldots, \Sigma_{i+1}) - E(f(\Sigma)|\Sigma_1, \ldots, \Sigma_i)| \leq c_i$$

Then

$$Pr(|f - E(f)| > t) \leq 2\exp\left(\frac{-t^2}{2\sum_i c_i^2}\right).$$

□

We will make use of Azuma's inequality by defining Σ_i to indicate the ith vertex to be exposed, for $i = 1, \ldots, \alpha$ and $f(\Sigma) = Z_\alpha$. We also define $E_{i+1}(x) = E(Z_\alpha|\Sigma_1, \ldots, \Sigma_{i+1})$, where Σ_{i+1} is the event that the $(i+1)th$ vertex is x. We would like to bound

$$|E(f(\Sigma)|\Sigma_1, \ldots, \Sigma_{i+1}) - E(f(\Sigma)|\Sigma_1, \ldots, \Sigma_i)|$$

We will do this by first bounding $|E_{i+1}(x) - E_{i+1}(y)|$ for any x, y. Let u, v be any two vertices. Suppose that we are choosing the $(i+1)$st vertex. We are therefore left with $n-i$ vertices. Note that by ignoring u, v, the distribution of the order in which the remaining vertices are exposed is unaffacted by the positions of u and v.

Let Ω be the set of the first remaining $\alpha - i - 3$ vertices in this order,

$$Z_\alpha = Z_i + \sum_{\Omega}(i + j - 2s - \frac{2}{3}t) + deg(y_1) - \big(2s + \frac{2}{3}t\big) + deg(y_2) - \big(2s\frac{2}{3}t\big).$$

where y_1 is either u or v and y_2 is either u, v or the next vertex in the order. Hence we see that the choice between u and v can only change Z_α by an amount equal to the maximum degree which is $w(n)$. Hence,

$$\max_{x,y}|E_{i+1}(x) - E_{i+1}(y)| \leq w(n).$$

Given the fact that

$$E(f(\Sigma)|\Sigma_1, \ldots, \Sigma_i) = \sum_x Pr(\text{x is chosen})\ E_{i+1}(x) \leq \max_x\ E_{i+1}(x),$$

we get that

$$\begin{aligned} &|E(f(\Sigma)|\Sigma_1, \ldots, \Sigma_{i+1}) - E(f(\Sigma)|\Sigma_1, \ldots, \Sigma_i)| \\ &\leq \max_{xy}\ \ |E_{i+1}(x) - E_{i+1}(y)| \leq w(n). \end{aligned}$$

Therefore, the probability that $Z_\alpha > -2$ is at most $Z_\alpha > 0$ which is

$$Pr(|Z_\alpha - E(Z_\alpha)|) > E(Z_\alpha).$$

By Azuma's inequality, this is at most

$$2\exp(-\frac{(Q/2\alpha)^2}{2\sum w(n)^2}) = 2\exp(-\frac{(Q/2\alpha)^2}{2\alpha w(n)^2})$$

Substituting $\alpha = Sw(n)^2 log(n)$, we get

$$2exp(-\frac{(Q^2/4Slog(n))}{2} = 2n^{-Q^2/8S} < n^{-2}.$$

The last inequality holds by substituting $S = \frac{17}{Q^2}$ and working through.

Hence the probability that a randomly chosen vertex lies on a component of size at least α is $o(n^{-1})$ and hence the expected number of such vertices is $o(1)$, so asymptotically none are expected to exist.

A.3.1 Very few configuration cycles

We we will now show that there is asymptotically no component with more than one configuration cycle, when the conditions of the second case of the theorem are satisfied. We will do this by showing that asymptotically we have very few back-edges.

Remark 17. *Looking at our algorithm, we see that X_r the number of open vertices decreases by at most 2 at every execution of step 2. Therefore by lemma (9), the size of any component is at most α implies that $X_r < 2\alpha$ at any step r during the execution of our algorithm. More precisely, at step r we must have that $X_r < 2(\alpha - r)$.*

Lemma 10. *Let G be a graph satisfying the conditions of the second case of the theorem. Then G almost sure has no components with 2 cycles.*

Proof. Fix any vertex v. We start our algorithm at step 1 with this vertex. We suppose that v lies in a component with more than one cycle. We will show that this happens with a very small probability and therefore asymptotically no such vertices are expected to es

Because at every iteration in our algorithm we either expose a new vertex or form a back-edge, we must have that the second back-edge is formed before $(\alpha + 2)$ steps or else we would have exposed more than α vertices which we saw by lemma (9) has a probability less than n^{-2}.

Suppose the first and second back edges are formed at step A and B respectively such that $0 \leq A \leq B \leq \alpha + 2$. The probability that we form two backedges is at most

$$\sum_{A=0}^{\alpha+1} \sum_{B=A}^{\alpha+2} \left(\frac{2(\alpha - A)}{Kn - 2(\alpha - A) - 3} \right) \left(\frac{2(\alpha - B)}{Kn - 2(\alpha - B) - 3} \right).$$

This is because the number of open vertices is less than $2(\alpha - A)$ or $2(\alpha - B)$ and the number of elements left in S is more than $Kn - 2(\alpha - A) - 3$ or $Kn - 2(\alpha - B) - 3$. This probability is at most

$$\begin{aligned}
&\sum_{A=0}^{\alpha+1} \sum_{B=A}^{\alpha+2} \left(\frac{2(\alpha - B)}{Kn - 2(\alpha - B) - 3} \right)^2 \\
&\leq \sum_{A=0}^{\alpha+2} \sum_{B=0}^{\alpha+2} \left(\frac{2(\alpha - B)}{Kn - 2(\alpha - B) - 3} \right)^2 \\
&\leq (\alpha + 2) \sum_{B=0}^{\alpha+2} \left(\frac{2(\alpha - B)}{Kn - 2(\alpha - B) - 3} \right)^2 \\
&\leq (\alpha + 2) \frac{1}{(Kn - 2(\alpha + 2) - 3)^2} \sum_{B=0}^{\alpha+2} (2(\alpha - B))^2 \\
&= (\alpha + 2) \frac{1}{(Kn - 2(\alpha + 2) - 3)^2} \sum_{B=0}^{\alpha+2} (2B)^2 \\
&\leq (\alpha + 2) \frac{4}{(Kn - 2(\alpha + 2))^2} (\alpha + 2)^3.
\end{aligned}$$

Because $\alpha = Sw(n)^2 log(n)$, if we take $w(n) = n^{1/8-\epsilon}$ for any $\epsilon > 0$, we get

$$\frac{1}{(Kn - 2(\alpha + 2))^2} (\alpha + 2)^4 = o(n^{-1})$$

So the probability that a random vertex v lies in a component with two cycles is at most $o(n^{-1})$. Therefore the expected number of these is $o(1)$. So with high probability there no components with more than once cycle as n tends to infinity.

□

A.4 A giant component

In this section, we will consider the first case of theorem (3). We will show that given that $D, Q > 0$, then with high probability our graph has a giant component and at least a linear number of cycles.

We will proceed as follows: First, we start our configuration building algorithm with any given vertex, and we show that after a certain number of step, Z_q very large with high probability. We will then use relation (A.2.2) to show that X_r is also very large with high probability. Having shown that the number of open vertices X_r is very large, we will deduce that that with high probability our configuration building algorithm will form a large number of back-edges and exposes a large number of new vertices, hence forming a giant component and a large number of cycles.

Lemma 11. *If $Q > 0$, then there exists $0 < \epsilon < 1$ and $0 < \Delta < min(\frac{\beta}{2}, \frac{K}{2})$ such that for all $0 < \delta < \Delta$, then a.s. $Z_{\delta n} > \epsilon \delta n$. The probability of the converse is z_1^n for some $0 < z_1 < 1$.*

Proof. In what follows, we will assume for simplicity that δn is an integer.

Recall that in the proof of lemma (9), we bounded the expected increase in Z_q after a number of steps of $\alpha = o(n)$. We then used this to have a bound on the expected value of Z_q itself. Then we showed using Azuma's inequality that with high probability we cannot deviate too far from the mean.

We will proceed similarly. However, the problem here is that we want to bound the expected increase in Z_q after a linear number of steps δn. This causes the probability of choosing a copy of a vertex of a certain degree to shift significantly.

To get around this, we will define a new variable Z_q^* that behaves much more nicely than Z_q such that Z_q majorises Z_q^*, i.e. that:

$$Pr(Z_q \geq x) \leq Pr(Z_q^* \geq x).$$

Define q_{st} to be the initial probability that that we choose a copy of a vertex of degree (s,t). We have that:

$$q_{i,j} = \frac{(i+j)d_{i,j}}{\sum_{i,j}(i+j)d_{i,j}} = (i+j)\frac{p_{i,j}}{K}.$$

This probability is likely to shift after δn steps. We define Z_q^* by fixing a number k^* and a sequence of probability values $\phi_{1,0}, \phi_{0,1}, \ldots, \phi_{i+j=k^*}$. Such that Z_q^* is the sum of all $(i+j-2s-\frac{3}{2}t)$ by the time the qth vertex is exposed, with the difference that every vertex of joint degree (i,j) is chosen with probability $\phi_{i,j}$ at every step, and that if we select a vertex of sum degree greater than k^* we treat as having sum degree 1 i.e. subtract $2s + \frac{3}{2}t$.

Clearly, if after q steps $q_{i,j} \geq \phi_{i,j}$ for $2 \leq i+j \leq k^*$, then:

$$Pr(Z_q \geq x) \leq Pr(Z_q^* \geq x).$$

for any x. Therefore, it suffices to find such k^* and $\phi_{i,j}$ such that after δn steps Z_q^* is at least $\epsilon\delta n$, this will be achieved by finding Z_q^* that has a positive expected increase.

Given that $Q > 0$, we have that:

$$\begin{aligned} Q &= \sum_{i,j}(i+j-2s-\frac{3}{2}t)\frac{(i+j)d_{i,j}}{\sum_{i,j}(i+j)d_{i,j}} = \sum_{i,j}(i+j-2s-\frac{3}{2}t)q_{i,j} \\ &= \sum_{i,j}(i+j-s-\frac{1}{2}t-1)q_{i,j} = \sum_{i,j}(i+j-s-1) - \sum_{i,j}(s+\frac{1}{2}t)q_{i,j} \\ &= \sum_{i+j\geq 2}(i+j-1)q_{i,j} - (s+\frac{1}{2}t) > 0 \end{aligned}$$

Because the asymptotic degree sequence is well behaved, see definition (31). We can find a k^* such that:

$$\sum_{i,j\geq 2}(i+j-1)q_{i,j} > (s+\frac{1}{2}t) + \epsilon'.$$

Therefore we can also find a sequence $\phi_{i,j}$ of joint probability values such that:

- $\phi_{i,j} < q_{i,j}, \quad \text{for} \quad 2 \leq (i+j) \leq k^*$.
- $\phi_{0,1} + \phi_{1,0} = \phi_{1,1} + \cdots + \phi_{i+j=k^*}$.
- $\sum_{i,j\geq 2}(i+j-1)\phi_{i,j} = (s+\frac{1}{2}t) + \frac{\epsilon'}{2}$.

This gives that:

$$\sum_{i,j\geq 2}(i+j-2s-\frac{3}{2}t)\phi_{i,j}=\frac{\epsilon'}{2}.$$

We construct such a joint probability sequence as follows: Given that $q_{i,j}=(i+j)\frac{p_{i,j}}{K}$, for $(i+j)\geq 2$, choose any $\Delta_{i,j}>0$ such that:

$$\phi_{i,j}\geq<(i+j-\Delta_{i,j})\frac{p_{i,j}}{K}<q_{i,j}$$

Taking $\Delta=min_{i,j}\{\Delta_{1,1},\Delta_{2,0}\dots\Delta_{i+j=k^*},\frac{\beta}{2},\frac{K}{2}\}$, then after exposing up to Δn vertices, the probability of choosing a copy of a vertex of sum degree $2\leq(i+j)\leq k^*$ is at least $\phi_{i,j}$. Therefore for $0\leq q\leq\Delta n$:

$$Pr(Z_q\geq x)\leq Pr(Z_q^*\geq x).$$

Let us now consider the variable Z_q^* with the following properties:

- $Z_0^*=0$.
- $Z_{q+1}^*=Z_q^*+(i+j-2s-\frac{3}{2}t)$, with probability $\phi_{i,j}$ for $2\leq(i+j)\leq k^*$.

This variables has expected increase $\frac{\epsilon'}{2}$ at every step q. Therefore, after δn steps, for $delta<Delta$, its expected value is $\frac{\epsilon'}{2}\delta n$. By Chernoff's inequality, see [8], we get that:

$$Pr(Z_{\delta n}^*\leq\frac{1}{2}E(Z_{\delta n}^*))\leq\exp\left(-\frac{E(Z_{\delta n}^*)}{4}\right)$$
$$Pr(Z_{\delta n}^*>\frac{\epsilon'}{4}\delta n)\geq 1-\exp\left(-\frac{\epsilon'\delta n}{8}\right)$$

Therefore, if take $\epsilon=\frac{\epsilon'}{4}$ we get that with high probability $Z_{\delta n}>\epsilon\delta n$.

□

Having shown that Z_q is very large for q large enough. We will now show that X_r also becomes very large at some point before Δn.

Lemma 12. *If $Q>0$ and Q finite, then there exists δ' such that for all $0<\delta\leq\delta'$, there a.s. exists $0<\eta<1$ such that $Z_{\delta n}\leq\eta n$ where $\eta=min\{\frac{sK}{4},\frac{tK}{4}\}$. The probability of the converse is at most $(z_2)^n$ for some $0<z_2<1$.*

Proof. The initial expected increase in Z_q is given by Q. We will show that even after δn steps this expected increase is not that large. We will the use an upper bound on the expected increase to bound $E(Z_q)$ and then Chernoff's inequality to bound Z_q itself.

The initial expected increase in Z_q is given by Q:

$$Q=\sum_{i,j}(i+j-2s-\frac{3t}{2})q_{i,j}$$
$$=\sum_{i,j}(i+j-2s-\frac{3t}{2})(i+j)\frac{p_{i,j}}{K}.$$

In the worst case, the first δn exposed vertices are all of sum degree $i+j=1$. After δn steps the probability of choosing a copy of a vertex of sum degree $i+j\geq 2$ is :

$$q_{i,j} = (i+j)\frac{p_{i,j}}{K-\delta} \leq \frac{p_{i,j}}{K/2} \leq 2q_{i,j}.$$

This implies that for all $q \leq \delta n$ the expected increase in Z_q is at most $2Q$. Therefore:

$$E(Z_q) \leq 2Q\delta n.$$

If we take $\delta \leq \min\{\frac{1}{2}\frac{sK}{4}, \frac{1}{2}\frac{tK}{4}\} = \delta'$, then by Chernoff's inequality:

$$Z_{\delta n} \leq \min\{\frac{sK}{4}, \frac{tK}{4}\}$$

with a probability exponential in n.

□

Corollary 3. *Using equation (A.2.2) we obtain that because $Z_{\delta n} \leq \eta n$ for any $0 < \delta \leq \delta'$, we must have that $X_{I_{\delta n} \leq \eta n}$.*

Lemma 13. *If $Q > 0$, then there exists $0 < \delta'' < \delta'$ such that for any $0 < \delta \leq \delta''$, there a.s. exist an R, $0 < R < R_{\delta n}$ such that $X_R > \gamma n$ where $\gamma = \frac{\epsilon\delta}{4}$. The probability of the converse is $(z_2)^n$ for some $0 < z_2 < 1$.*

Proof. We will bound X_r using relation (A.2.2):

$$\begin{aligned} Z_q &\leq X_{R_q} + 2Y_{R_q} + \frac{3}{2}Y'_{R_q} - \frac{3}{2} \\ Z_q &\leq X_{R_q} + 2Y_{R_q} + \frac{3}{2}Y'_{R_q} \\ X_{R_q} &\geq Z_q - 2Y_{R_q} - \frac{3}{2}Y'_{R_q} \\ X_{R_q} &\geq Z_q - 2(Y_{R_q} + Y'_{R_q}). \end{aligned} \tag{A.4.1}$$

We also have that because $X_r \geq 0$:

$$\begin{aligned} Z_q &\geq 2Y_{R_q} + \frac{3}{2}Y'_{R_q} - 2 \\ &\geq \frac{3}{2}(Y_{R_q} + Y'_{R_q}) - 2 \end{aligned} \tag{A.4.2}$$

Therefore if we want to bound X_{R_q}, we will have to bound the number of back edges formed $(Y_{R_q} + Y'_{R_q})$ from above. We will do this by counting the number of back edges formed before $X_r > \gamma n$, or $R_{\delta n}$ pairs have been formed.

At any step r, $1 \leq r < R_{\delta n}$, the probability that we form a back edge is the probability of choosing an open vertex in step 2 of algorithm (2), which is at most $\frac{X_r}{Kn-2r}$. Let us now bound R_q the number of steps required to expose q vertices:

$$R_q \leq q + Y_{R_q} + Y'_{R_q} \leq q + \frac{2}{3}Z_q + \frac{4}{3}$$

Using the result of lemma (12), we have that $Z_q \leq \frac{K}{4}$. This gives:

$$R_q \leq q + \frac{K}{6}n + \frac{4}{3} \leq \delta n + \frac{K}{6}n$$

for n large enough. The probability p of forming a back-edge when $X_r \leq \gamma n$ is at most:

$$p = \frac{X_r}{Kn - 2\frac{K}{6}n - 2\delta n} \leq \frac{\epsilon\delta/16}{\frac{2K}{3} - 2\delta}$$

Consequently, the number of back edges formed has an expected value of at most:

$$E(Y_{R_q} + Y'_{R_q}) \leq pI_{\delta n} \leq \frac{\epsilon\delta/16}{\frac{2K}{3} - 2\delta}(\delta + \frac{K}{6})n$$

We would like this expected value to be less than $\frac{\epsilon\delta}{8}n$ i.e. :

$$\begin{aligned}
\frac{\epsilon\delta/4}{\frac{2K}{3} - 2\delta}(\delta + \frac{K}{6})n &< \frac{\epsilon\delta}{8}n \\
\frac{\delta + \frac{K}{6}}{\frac{2K}{6} - \delta} &< \frac{1}{2} \\
\delta + \frac{K}{6} &< \frac{K}{3} - \delta \\
2\delta &< \frac{K}{6}
\end{aligned}$$

If we take $\delta < \frac{K}{12} = \delta''$, we get by Chernoff's inequality:

$$Pr(Y_{R_q} + Y'_{R_q} > \frac{\epsilon\delta}{4}n) \leq (z_3)^n$$

for some $0 < z_3 < 1$.

Therefore, if for all $1 \leq r < R_{\delta n}$, $X_r \leq \frac{\epsilon\delta}{4}$, we get that with high probability:

$$Y_{R_q} + Y'_{R_q} \leq \frac{\epsilon\delta}{4}n.$$

Using inequality (A.4.1) we obtain

$$\begin{aligned}
X_{R_{\delta n}} &\geq Z_{\delta n} - 2(Y_{R_{\delta n}} + Y'_{R_{\delta n}}) \\
&\geq \epsilon\delta - \frac{\epsilon\delta}{2} = \frac{\epsilon\delta}{2} > \frac{\epsilon\delta}{4}.
\end{aligned}$$

□

Lemma 14. *If $Q > 0$, there exists constants c_1, c_2 such that the component being exposed at step $R \leq R_{\delta'' n}$ has a.s. at least $c_1 n$ vertices and c_2 cycles. The probability of the converse is $(z_4)^n$ for some $0 < z_4 < 1$.*

Proof. We have shown that there exists a step $R \leq \delta'' n$ such that $X_R > \gamma n$, $0 < \gamma < 1$. We will show with high probability $c_1 n$ of the X_R open copies will be matched with unexposed vertices and that $c_2 n$ will be matched with other open copies.

Among the X_R open copies at step I, at least half of these must be of the same colour $\mathcal{C}$. We construct a set B containing all open copies of colour $\mathcal{C}$. This set has size at least $\frac{\gamma}{2}n$.

Also, qt step $I_{\delta n}$, we have only exposed δn vertices. From lemma (11), δ is at most $\frac{\beta}{2}$ this implies there is at least $\frac{\beta}{2}$ remaining vertices that have copies of both colours. We create a set A containing one copy of each of these vertices with colour $\mathcal{C}$.

After I steps there at least $(Kn - 2R)$ copies left to be matched. We will show that $c_1 n$ copies will be matched with members of A and $c_2 n$ copies will be matched with members of B. Our configuration building algorithm pairs up, these $(Kn - 2R)$ open copies of vertices uniformly. It essentially creates $(Kn - 2R)/2$ pairs, with each pair created with an equal probability.

In general, given any two sets A and B which are subsets of a set C , the probability we create a pair containing a copy from A and a copy from B from a set C is

$$\frac{|A||B|}{\binom{|C|}{2}}$$

where $|A|$ denotes the size of set A, and the expected number of these is:

$$\frac{|A|\ |B|}{\binom{|C|}{2}}\ \frac{|C|}{2} \leq \frac{|A|\ |B|}{|C|}.$$

Therefore, the expected number of pairs containing one copy from A and one copy from B in our configuration is:

$$\leq \frac{\beta/2\ \gamma/2\ n}{Kn - 2R} \leq \frac{\beta/2\gamma/2n}{Kn - 2\delta n - \frac{K}{3}n} = 2c_1 n + o(n).$$

for some constant $c_1 > 0$. The expected number of pairs containing two copies of B is:

$$\leq \frac{\gamma/2n\ \gamma/2n}{Kn - 2R} \leq \frac{\gamma/2n\ \gamma/2n}{Kn - 2\delta n - \frac{K}{3}n} = 2c_2 n + o(n).$$

for some constant $c_2 > 0$. Finally, using Chernoff's inequality we get that the number of such pairs is less than half their expected values with a probability $(z_4)^n$ for some $0 < z_4 < 1$. So, with high probability we have at least $c_1 n$ vertices in our component being exposed and at least $c_2 n$ back-edges.

□

Lemma 15. *Given a configuration F as described in the first case of theorem (3), then F a.s. has at most one component with more than $T\log(n)$ vertices for an appropriate choice of constant T.*

Proof. We have already shown that F has at least one component of size $c_1 n$ for some constant $c_1 > 0$. We have also shown that there exist an R, $R \leq R_{c_1 n}$ such that $X_R > \gamma n$ where $\gamma = \min \frac{\epsilon c_1}{4}, \delta''$.

We will look at pairs of vertices (u, v) and show that the probability that u and v belong to different components of size at least $c_1 n$ and $T\log(n)$ respectively. We call these components C_1 and C_2 respectively. We will show that this happens with probability $o(n^{-2})$ and therefore the expected number of such pairs is zero.

We suppose that such a pair exists and we start algorithm (2) with any copy of vertex u. If after R steps of algorithm (2), we are no longer exposing C_1 then u does not lie on a component of size at $c_1 n$, and if we have exposed a copy of v then u and v are in the same component so we will assume neither event happens.

We modify our exploration algorithm slightly, by stopping the exploration of C_1 after R steps, and starting to explore v's component. This is legitimate because u and v are in different components and this still produces a random configuration.

We will show that with high probability one of the vertices of C_2 will be matched with one of the X_R open copies of the exploration of C_1. Since $X_r > \gamma n$), and the number of available copies to be matched with at any point during the exploration of C_2 is at most Kn, we get that the probability of choosing one the X_r open copies during the exploration of $C-2$ is at least: $\frac{\gamma}{K}$.

Because C_2 has at least $T\log(n)$ vertices. The probability of matching a vertex of C_2 with one of the X_R open copies from the open exploration of C_1 is at most

$$\left(1 - \frac{\gamma}{K}\right)^{T\log(n)} = \left(e^{-c}\right)^{T\log(n)}.$$

for some constant c. Taking $T > 2c$ give:

$$\left(1 - \frac{\gamma}{K}\right)^{T\log(n)} = o(n^{-2}).$$

Therefore the expected number of pairs (u, v) that lie on components of size at $c_1 n$ and $T\log(n)$ respectively is $o(1)$, so a.s. none exist.

□

Lemma 16. *Given a configuration F as described in the conditions of the first case of theorem (3), then F a.s. no components of size at most $T\log(n)$ wih more than one cycle.*

Proof. We have shown that a configuration F satisfying conditions of the first case of theorem (3 has a.s. exactly one component of size at least $c_1 n$ for some $0 < c_1 < 1$, and that all other components have size at most $T\log(n)$.

Suppose there exists one such a component with at least two cycles. Let v be a vertex in such a component. We start algorithm (2) at vertex v. We will show that the probability of having two back edges $o(n^{-1}$ and therefore no such vertices are expected to exist.

Because the size if the component of v is at most $T\log(n)$, each vertex in it has degree at most $T\log(n)$ as well. We therefore have that $X_r \leq T^2\log(n)^2$ at any step r, because the maximum number of copies of vertices of this component it at most the number of vertices times the maximum degree. For the same reason we have that the component is entirely exposed in at most $T^2\log(n)^2$ steps.

The probability that a back edge is formed at any step r is at most:

$$\frac{X_r}{Kn-2r} \leq \frac{T^2\log(n)^2}{Kn-2r} \leq \frac{T^2\log(n)^2}{Kn-2T^2\log(n)^2} = o(n^{-1/2}).$$

The probability of forming at least 2 back edges is at most:

$$\binom{T\log(n)}{2}(n^{-1/2})^2 = o(n^{-1}).$$

Therefore, the expected number of vertices in components of size at most $T\log(n)$ and more than one cycle is o(1) and therefore a.s. none exist.

□

Bibliography

[1] A.L.Barabási and R.Albert, *Emergence of scaling in random networks*, Science **286**, 509-512 (1999).

[2] E.A.Bender, E.R.Canfield, *The asymptotic number of labeled graphs with given degree sequences*, Journal of combinatorial theory A **24**, 296-307 (1978).

[3] B.Bollobás, *Random Graphs*, Academic Press, New York (1985).

[4] S.Boccaletti, V.Latora, Y.Moreno, M.Chavez, D.Hwang. *Complex networks: Structure and dynamics*, Physics Reports , vol. 424, no. 4-5, 175-308 (2006).

[5] B.Bollobás, S.Janson and O.Riordan, *Sparse Random Graphs with Clustering*, preprint 2008, arXiv:0807.2040.

[6] B.Bollobásás, O.Riordan, *The diameter of a scale free random graph*, Combinatorica **34**, 5-34 (2002).

[7] B.Bollobás, O.Riordan, J.Spencer, G.Tusnády, *The degree sequence of a scale-free random graph process*, Random structures and algorithms **18**, 279-290 (2001).

[8] W.Feller, *An introduction to probability theory and its applications*, Vol 1, Wiley (1966).

[9] F.Chung, L.Lu, *The average distance in random graphs with given expected degrees*, Internet Mathematics **1**, 15879-15882 (2002).

[10] R.Cohen, K.Erez, D.Ben-Avraham, S.Havlin1, *Resilience of the Internet to Random Breakdowns*, Phys. Rev. Lett. 85, 4626 - 4628 (2000).

[11] S.N.Dorogovtsev, J.F.F.Mendes, A.N.Samukhin, *Metric structure of random networks*, Nucl. Phys. B 653, 307 (2003).

[12] P.Erdős and A.Rényi, *On Random graphs*, publicationes mathematicae **6**, 290-297 (1959).

[13] A.Fronczak, J.A.Holyst, M.Jedynak, J. Sienkiewicz, *Higher order clustering coefficients in Barabasi-Albert networks*, Physica A **316**, 688-694 (2002).

[14] J.P.Gleeson, *Bond percolation on a class of clustered random graphs*, preprint (2009), arXiv:0904:4292.

[15] G.Grimmett, *Percolation*, Springer (1999).

[16] P.Holme , B.J.Kim, *Growing scale free networks with tunable clustering*, Phys. Rev. E **65** 036133 (2005).

[17] S. Janson, *The probability that a random multigraph is simple*. Combinatorics, Probability and Computing **18**, 205-225, 2009.

[18] S.Janson, M.Luczak, *A new approach to the giant component*, Random Structures and Algorithms **34**, 197-216 (2009).

[19] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.-L. Barabs. *The large-scale organization of metabolic networks*. Nature 407 651 (2000).

[20] B. Jiang, C. Claramunt, *Topological Analysis of Urban Street Networks*, Environ. Plann. B 31 151 (2004).

[21] S.Milgram , *The small world problem*, psychology today **2** 60-67 (1967).

[22] M.Molloy and B.Reed, *A critical point for random graphs with a given degree sequence*, Random structures and algorithms **6** 161-179 (1995).

[23] M.Molloy and B.Reed, *The size of the giant component of a random graph with a given degree sequence*, Combinatorics probability and computing **7**, 295-305 (1998).

[24] M.E.J.Newman, S.H.Strogatz, D.J.Watts, *Random graphs with arbitrary degree distributions and their applications*, Phys. Rev E **64**, 026118 (2001).

[25] M. E. J. Newman, *The structure and function of complex networks*, SIAM Review **45**, 167256 (2003).

[26] M.E.J.Newman, S.H.Strogatz, D.J.Watts, *Random graphs with arbitrary degree distributions and their applications*, Phys. Rev. E 64, 026118 (2001).

[27] M.E.J.Newman, *The structure and function of complex networks*, Phys. Rev. E 67, 026126 (2003).

[28] M.E.J.Newman, Juyong Park. Phys. Rev. E 68, 036122 (2003).

[29] M.E.J.Newman, *Random graphs with clustering*, Physical review letters, **103**, 058701 (2009).

[30] M. E. J. Newman, *in Handbook of Graphs and Networks*, S. Bornholdt and H. G. Schuster (eds.), Wiley-VCH, Berlin (2003)

[31] M.A.Porter, P.J.Mucha, , M.E.J.Newman, A.J.Friend. *Community Structure in the United States House of Representatives*. Physica A, Vol. 386, No. 1: 414-438 (2007).

[32] A.L.Traud, E.D.Kelsic, P.J.Mucha, M.A.Porter. *Community Structure in Online Collegiate Social Networks*. Submitted to SIAM Review, arXiv:0809.0690 (2008).

[33] D.Price, *A general theory of bibliometric and other cumulative advantage processes*, Journal of the American Society for Information Science **27** 292-307 (1976).

[34] D.J.Watts, S.H.Strogatz, *Collective dynamics of small world networks*, Nature **393**, 440-442 (1998).

[35] T.Yu-Song, A.O.Sousa, K.Ling-Jiang, L.Mu-Ren, *Combined update scheme in the Sznajd model*, Physica A, Volume 370, Issue 2, 727-733 (2006).

[36] S. R.Broadbent, J.M.Hammersley *Percolation processes. I. Crystals and Mazes*, Proc Camb Philos Soc, vol. 53, no. 3, pp. 629-641 (1957).